Proteomic Analysis of the *Clostridium thermocellum* Cellulosome

Nicholas Gold

A Thesis

in

The Department

of

Biology

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Biology) at
Concordia University
Montréal, Québec, Canada

November 2007

Your file Votre référence
ISBN: 978-0-494-40852-0
Our file Notre référence
ISBN: 978-0-494-40852-0

ABSTRACT

Proteomic Analysis of the *Clostridium thermocellum* Cellulosome

Nicholas Gold

A metabolic isotope-labelling strategy was used in conjunction with nanoLC-ESI-MS peptide sequencing to assess quantitative alterations in the expression patterns of subunits within cellulosomes of the cellulolytic bacterium *Clostridium thermocellum*, grown on either cellulose or cellobiose. The effects of adding xylan, pectin and galactomannan to these cultures were also explored. In total, 55 cellulosomal proteins were detected, including 50 type I dockerin-containing proteins, which count among them all but one of the known docking components and 28 new subunits. All differential expression data was normalized to scaffoldin CipA such that protein-per-cellulosome was compared for growth between the different substrates. Proteins that exhibited higher expression in cellulosomes from cellulose-grown cells as compared to cellobiose-grown cells were: cell-surface anchor protein OlpB; exoglucanases CelS and CelK; and GH9 endoglucanase CelJ. Conversely, lower expression in cellulosomes from cells grown on cellulose as compared to cellobiose was observed for GH8 endoglucanase CelA; GH5 endoglucanases CelB, CelE, CelG; and hemicellulases XynA, XynC, XynZ, XghA. GH9 cellulases were the most abundant group of enzymes per CipA when cells were grown on cellulose, while hemicellulases were the most abundant group on cellobiose. The results support the existing theory that expression of scaffoldin-related proteins is coordinately regulated by a catabolite repression type of mechanism, as well as the prior observation that xylanase expression is subject to a growth rate-independent type of regulation. However, concerning transcriptional control of cellulases, which had also been previously shown to be subject to catabolite repression, a novel distinction was observed with respect to endoglucanases.

ACKNOWLEDGEMENTS

I want to thank Dr. Vincent Martin for giving me the opportunity to prove myself, for trusting me to handle his expensive machines, and for taking the time to show me how it's all done. I am grateful for his guidance, his support, and his friendship.

I am grateful to Dr. Justin Powlowski and Dr. Reginald Storms for sitting on my committee. I also thank Dr. Storms and Dr. Emma Master for helping to review the published manuscript that is the core of this thesis.

I also want to thank my lab mates whose camaraderie made daily work fun and productive. Kane LaRue helped in the development of the minimal medium we used. Euan Burton helped in the daily operation of the LC-MS.

Thanks to Rebecca Sydenham, Yun Zheng and Donald Patton for their help in setting up the enzymatic assays.

My parents and family get innumerable thanks of course for seeing me through this part of my life, as does my best friend and number one girl Klara.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ABC	adenosine-binding cassette
ACN	acetonitrile
BCA	bicinchoninic acid
CBD#	cellulose binding domain family #
CBP	consolidated bioprocessing
CE#	carbohydrate esterase family #
CID	collision induced dissociation
CMC	carboxymethyl cellulose
Coh1	type I cohesin domain
Coh2	type II cohesin domain
Doc1	type I dockerin domain
Doc2	type II dockerin domain
DOE	U.S. Department of Energy
DTT	dithiothreitol
emPAI	exponentially modified protein abundance index
ESI	electrospray ionization
FA	formic acid
Fn	fibronectin
GH#	glycoside hydrolase family #
GHG	greenhouse gas
Ig	immunoglobulin
JGI	Joint Genome Institute
nanoLC	nanovolume liquid chromatography
MS	mass spectrometry or mass spectrometer
MS/MS	tandem MS
MWCO	molecular weight cut-off
m/z	mass to charge ratio
NCBI	National Center for Biotechnology Information
PAGE	polyacrylamide gel electrophoresis
PASC	phosphoric acid swollen cellulose
PNP/PNPC	*p*-nitrophenol/*p*-nitrophenyl-β-D-cellobioside
P_{pro}	protein probability
PQD	pulsed-Q dissociation
PTM	post-translational modification
SDS	sodium dodecyl sulfate
SD	standard deviation
SE	standard error
SLH	surface-layer homology
S/N	signal to noise ratio
SSF	simultaneous saccharification and fermentation
TPCK	*N*-tosyl-L-phenylalanine chloromethyl ketone
XC	cross correlation
XIC	extracted ion chromatogram
XPM	xylan + pectin + locust bean gum

1.1. OBJECTIVES

In this study of cellulosomal gene expression at the proteome level in *Clostridium thermocellum*, there were two main objectives: first, to query the composition of the cellulosome protein complex using nanoLC-ESI-MS peptide sequencing; and, second, to quantitatively assess changes in the subunit profiles within cellulosomes isolated from cells grown on Avicel (microcrystalline cellulose) versus cellobiose as carbon source. The addition of hemicelluloses to these substrates was also investigated. Quantitation was achieved using a metabolic isotope-labelling strategy in conjunction with nanoLC-ESI-MS; a peptide counting technique was also applied to approximate the relative abundance of each cellulosome component per sample. In comparing cellulosomes from cells grown on different substrates, we expected to detect several novel gene products and also to uncover differences in protein expression that can shed more light on our understanding of the regulation of cellulosomal cellulases and hemicellulases. Cells grown on Avicel were expected to produce cellulosomes with increased levels of key enzymes for degradation of crystalline cellulose such as the processive exoglucanase CelS [1, 2].

1.2. STRUCTURE OF THE THESIS

The text is organized in the style of a journal article. The introduction presents background on *C. thermocellum* and lignocellulosic ethanol, as well as concepts in quantitative proteomics using mass spectrometry. Experimental techniques used are described in the methods section. Results section 4.2 describes published data for the comparison of cellulosomes from cells grown on Avicel versus cellobiose [3]. Section 4.3 describes data from the comparison of growth on Avicel and cellobiose with or without

hemicelluloses added. A discussion of these results concludes the main body of the text. Some information deemed peripheral to the essence of the thesis was deferred to appendices, although it is relevant to anyone wishing to take up the continuation of this work.

2. INTRODUCTION

2.1.1. Cellulosic ethanol

Research into the conversion of lignocellulosic biomass is driven by the pursuit of environmental security as well as energy security. Global climate change can have a range of significant impacts on extreme weather events, natural ecosystems, human health and economic activity [4]. There is general agreement in the scientific community that the rise in global temperatures is due to emissions of greenhouse gases like carbon dioxide, the main source of which is the burning of fossil fuels [4]. Worldwide levels of carbon dioxide emissions from fossil fuels increased by a record of 4.5% in 2004, to 7.57 billion tons of carbon [5]. In 2005, Canada's total greenhouse gas (GHG) emissions were estimated at 747 megatons of carbon dioxide equivalent (up 25% from 1990), and energy production and consumption contributed about 82% of this [6]. In such a way, the problem of climate change is intertwined with the matter of energy supply, production and consumption. Solutions to climate change that seek to curb carbon dioxide emissions will thus also need to pose an alternative to fossil fuel as a source of energy. In view of

the projected shortages and increasing prices of fossil fuels, this is the perfect time for world economies to seek out alternative energy sources that can both reduce net GHG emissions and help alleviate their dependence on oil.

The international community has recognized the need to curb GHG emissions and has set goals, legally binding obligations to be sure, for doing so, in the form of the Kyoto Protocol, ratified by 141 countries in 2005. *Canada's Fourth National Report on Climate Change* describes federal and provincial initiatives being implemented to encourage 'cleaner' living and increased consideration for bioenergy technologies at the consumer level and by research and industry [7].

Bioenergy sources are of many different types, from hydrogen to biodiesel, bioethanol and biogas. Ethanol, as a high performance fuel for spark-ignition internal combustion, contains about two-thirds the energy per volume of gasoline, and can be used by automobiles in a blend with gasoline up to 20% ethanol with no modifications to the engine [8]. Fuel-flexible vehicles (FFVs) are capable of utilizing blends with up to 85% ethanol (E85). In Canada, 7% of all gasoline currently sold is blended with ethanol, and 11 new plant projects are projected to produce an additional 1.2 billion litres of ethanol by the end of this year [7]. Fuel ethanol is mass-produced from sugarcane in Brazil, where the energy produced powers the production process [9]. In the United States, large-scale fuel ethanol is made from starches in grains (corn, wheat, barley, rye), however the production processes are presently powered mostly by fossil fuels such that the net GHG emissions are not much lower than they are for gasoline [10]. What makes bioethanol attractive as a solution to GHG emission reduction is the fact that carbon dioxide exhausted by its combustion is offset by the carbon dioxide fixed during

photosynthetic growth of the feedstock [11, 12]. In theory, bioethanol production and consumption is thus considered a GHG neutral process, however the extent to which this zero net GHG cycle is maintained in practice depends on the fossil fuel inputs required for feedstock production, conversion, and utilization.

Ethanol as well as other fuels derived from cellulosic materials in plant cell walls has perhaps the greatest potential for reducing GHG levels while bringing about energy self-sufficiency [10, 13]. Cellulose is the most abundant organic polymer on Earth. Feedstocks for lignocellulosic ethanol are relatively inexpensive and can vary from dedicated energy crops (perennial grasses such as switchgrass and miscanthus) to agricultural plant wastes (corn stover, cereal straws, sugarcane bagasse) to industrial plant wastes (paper pulp, sawdust, wood chips) to municipal solid wastes [13, 14]. In production designs, lignin, a by-product of the biomass conversion process, can be burned instead of fossil fuel to power production. Thus, because the fossil inputs are low, the ratio of energy output to fossil energy input is high, and by corollary net GHG emissions are low as well, exceptionally so given the sheer abundance of carbon dioxide-fixing feedstock that is taken into the equation.

Lignocellulosic ethanol will create jobs and stimulate agriculture in regions incapable of supporting food crops. For the time being, however, slowed down by the once prohibitive cost of converting biomass into fermentable sugars, it remains on the cusp of being produced at the commercial scale. While pilot-scale (producing less than 1 million gallons of ethanol per year, MMgy) and demonstration-scale (1-10 MMgy) cellulosic ethanol plants are presently operational in Canada (Iogen Corporation; SunOpta BioProcess Inc.), the U.S. (National Renewable Energy Laboratory/Abengoa

Bioenergy Research & Development), Spain (Abengoa Bioenergy), Sweden (Etek), Denmark (Elsam), and the People's Republic of China (SunOpta BioProcess Inc.), the world's first commercial-scale biorefinery (10 MMgy) is scheduled to open in Canada by the end of 2007 (SunOpta BioProcess Inc./Greenfield Ethanol), and others could follow in the U.S. (Mascoma; Abengoa Bioenergy), the Netherlands (Nedalco), and the People's Republic of China (SunOpta BioProcess Inc.) by 2009 [15-18].

2.1.2. Strategies for overcoming the recalcitrance of cellulosic biomass

Conversion of lignocellulosic biomass involves two major steps: first, transformation of biomass into a utilizable carbon source; second, microbial fermentation of the resulting carbon to ethanol (or another valuable carbon-based chemical). While lignocellulosic ethanol technology is rapidly developing with the help of biotechnology, one of the main stumbling blocks to its economic production has been overcoming the recalcitrance of cellulosic materials to release their fermentable carbon.

The recalcitrance of cellulosic biomass resides in its heterogeneous composition and in the crystalline structure of cellulose. Linear chains of (up to 15,000) β-1,4 linked anhydrous glucosyl residues hydrogen bond to form tightly packed cellulose microfibrils. The tight packing, responsible for the crystallinity of cellulose, limits penetration of small molecules and cellulolytic enzymes [12]. In such a way, cellulose is highly resistant to hydrolysis, although crystallinity exists in varying degrees depending on the feedstock. Further complicating matters, cellulosic microfibrils are locked into a matrix with other structural biopolymers; hemicellulose tethers cellulosic microfibrils together as well as to lignin. By dry weight, the secondary cell walls of plants are composed of 38-50%

cellulose, 23-32% hemicellulose, and 15-25% lignin [19]. Lignin is a large, cross-linked macromolecule consisting of various types of substructures, organized in an apparently haphazard manner and incorporating three monolignol monomers, methoxylated to various degrees: *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol [20]. Removing lignin is crucial to the conversion process. As mentioned above, it can be recovered and burned to help power the process. Hemicellulose, on the other hand, represents another valuable cache of utilizable sugars in biomass, and one much more susceptible to hydrolysis due to its structure [19]. In contrast with cellulose, it is a branched polymer of up to only 200 subunits that can consist of many different sugar monomers besides glucose: hexoses galactose and rhamnose, as well as pentoses xylose (the most common), mannose, and arabinose [21].

Cellulosic biomass can in fact be separated and converted into its composite carbon in several ways. Gasification transforms lignocellulosic materials into gaseous carbon monoxide and hydrogen, which can be fermented to ethanol by some anaerobic bacteria (like *Clostridium ljungdahlii*) [22]. Alternatively, the raw materials can be broken down into sugars for subsequent fermentation by robust ethanologenic microorganisms that can utilize hexoses (traditionally *Saccharomyces cerevisiae*) and preferably pentoses as well. The saccharification step can be achieved in two ways. The first involves acid hydrolysis, which is expensive and can generate degradation products (like furfural and hydroxymethyl furfural) that are toxic to fermentation [23]. Better and cleaner hydrolysate yields can be obtained from a second method that calls for a pre-treatment step (by dilute acid, organic solvents, steam explosion, or ammonia fibre expansion) to remove lignin (and sometimes hemicellulose), followed by enzymatic

hydrolysis of the remaining cellulose (and hemicellulose) [24]. This second avenue offers more possibility for cost reduction and improvement via biotechnology. Enzymatic hydrolysis is currently carried out using cocktails of purified cellulolytic enzymes, patterned after fungal cell-free cellulase systems (such as that of *Trichoderma reesei*) and genetically modified to achieve optimal and synergistic hydrolysis of cellulose. While the prohibitive cost of these purified enzymes has been one of the key obstacles to economic production of cellulosic ethanol, biotechnological advances are driving down these costs. One of the major steps forward was taken in 2004-2005 when both Genencor International and Novozymes Inc., two enzyme producing companies commissioned by the U.S. National Renewable Energy Laboratory, achieved 30-fold reductions in overall enzyme costs, lowering the enzyme cost of ethanol production from around $5.00 to less than $0.20 per gallon [25].

Hydrolytic enzymes can be implemented with a fermenting microorganism in the same vessel for simultaneous saccharification and fermentation (SSF). Conversion solutions seeking to circumvent the cost of enzyme production implicate the cellulolytic microbes themselves. One strategy is the co-culturing of two or more 'specialist' microorganisms; the first being a specialist in cellulose hydrolysis, the second in hexose fermentation, and perhaps a third in pentose fermentation [26]. Another strategy termed consolidated bioprocessing (CBP) involves the genetic engineering of a single microorganism to accomplish all steps of the conversion process by itself: production of saccharolytic enzymes, hydrolysis of pretreated biomass to sugars, fermentation of hexoses, and fermentation of pentoses [12, 27]. One approach to CBP is to take cellulase and hemicellulase genes and transform them into a classic hexose fermentor like

Saccharomyces cerevisiae; another approach is to streamline a cellulolytic organism for industrial ethanol production.

2.2. *Clostridium thermocellum*

One microorganism receiving considerable attention for CBP implementations is the cellulolytic, ethanologenic, anaerobic, thermophilic Gram-positive bacterium *Clostridium thermocellum* [12, 26]. The reason for the great interest in *C. thermocellum* is that it has an exceptionally high hydrolysis rate against crystalline cellulose, exhibiting about 50-fold higher specific activity than *Trichoderma reesei* [26], one of the aerobic fungi traditionally drawn on for most large-scale conversion technologies [12]. Indeed, *C. thermocellum* is capable of solubilizing lignocellulosic materials like dilute-acid pre-treated mixed hardwoods [28]. It further utilizes the cellulose hydrolysates yielding ethanol, acetic acid, lactic acid, hydrogen gas and carbon dioxide as fermentation end-products [29]. Thus, it has the potential for CBP of cellulose to ethanol.

The thermophilic and anaerobic features of its nature also pose advantages to using *C. thermocellum* for large-scale ethanol fermentation from biomass, as enumerated by Demain et al. [26]. Thermophiles tend to be robust microorganisms with stable enzymes. Fermentation at high temperature would reduce the cost of cooling, be less prone to contamination, and facilitate removal and recovery of ethanol, thus reducing the requirement for a strain with high tolerance to ethanol. Anaerobes tend to have low cell growth yields and thus convert most of their substrate to product. Anaerobiosis would eliminate the cost of aeration in the fermentation tanks.

C. thermocellum grows readily on cellulose, cellobiose (the β-1,4-linked glucose dimer and repeating unit of cellulose), and laminaribiose (the β-1,3-linked glucose dimer), and after a lag on fructose, sucrose and glucose [26]. However, it cannot grow on pentoses like xylose even though it is capable of solubilizing hemicellulose such as xylan [30] and it has intracellular β-xylosidase activity [31]. Improving substrate utilization via a co-culture strategy is a possibility that would involve anaerobic thermophiles such as *Clostridium thermosaccharolyticum* [32] and *Clostridium thermohydrosulfuricum* [33], which are capable of metabolizing pentoses. Maximizing ethanol yield by eliminating the metabolic pathways leading to lactic acid and acetic acid production is also possible via genetic manipulation. The genome of *C. thermocellum* has been sequenced by the Joint Genome Institute (JGI). A genetic electrotransformation system has been developed for *C. thermocellum* specifically [34], and a knockout system for *Clostridia* (ClosTron) has also been established [35]. Efforts to raise ethanol tolerance are also being made [36]. Microarray technology is now available for wide-scale gene expression studies in *C. thermocellum* at the transcriptome level [37].

2.2.1. The *C. thermocellum* cellulase system

Aerobic cellulolytic organisms produce extracellular cell-free cellulases in high concentration. On the other hand, *C. thermocellum*, being an anaerobic cellulolytic bacterium that relies on ATP from glycolysis for cellular energy, cannot afford to produce large amounts of extracellular cellulase. Instead, it organizes its cellulolytic enzymes into highly efficient cell surface-bound protein complexes termed cellulosomes. Cellulosome complexes have also been observed in other bacteria like *Clostridium*

cellulovorans [38], *Clostridium cellulolyticum* [39], *Clostridium josui* [40], *Clostridium acetobutylicum* [41], *Acetovibrio cellulolyticus* [42], *Bacteroides cellulosolvens*, *Ruminococcus albus* [43], *Ruminococcus flavefaciens* [44], *Vibrio* sp., and the anaerobic fungal genera *Neocallimastix*, *Piromyces*, and *Orpinomyces* [45].

The *C. thermocellum* cellulase system comprises both cellulosomal and noncellulosomal cell-surface bound enzymes, although the latter are responsible for no more than 5% of the overall endoglucanase activity [46]. There are exo- and endo-β-1,4-glucanases, xylanases and other hemicellulases, and carbohydrate esterases. The presence of these different enzymatic activities (cellulolytic and hemicellulolytic) in the cellulosome is one reason *C. thermocellum* is so effective at overcoming the heterogeneity of plant cell wall materials [47, 48]. The high efficiency of the cellulosome is also attributed to the presence in optimal stoichiometry of catalytic domains that complement one another resulting in synergism, the phenomenon whereby certain combinations of enzymes (endo- and exoglucanase pairs; pairs of exoglucanases that process cellulose chains from reducing and non-reducing ends) collectively exhibit higher overall activity than the sum of their individual activities [12]. Synergistic action among enzymes within the cellulosome setting is further enhanced by cellulose targeting via the cellulose-binding domain of the complex's central structural protein, and also by appropriate spacing between individual catalytic subunits (for optimal channelling of substrate between them) [49]. Preferred proximity relationships between specific catalytic domains also appear to be possible contributors to the synergistic effect [50]. The tethering of enzymes within the cellulosome prevents their cooperativity from being hindered by steric interactions between free subunits [51]. The phenomenon of enzyme-

enzyme synergy exists for cell-free cellulase systems [52] as well as for complexed ones; however, it is thought to be more pronounced in the cellulosome. In a recent study, cell-free versions of bacterial enzymes as well as dockable chimeric fungal enzymes were created, and then the activities of cell-free enzyme pairs were compared with enzyme pairs docked onto a chimeric scaffoldin [53]; and synergy was observed for the right combination of complementary docked enzymes.

The *C. thermocellum* cellulosome is tethered to the cell surface, which means that the products of hydrolysis are also in proximity to the cell, where they can be taken up via adenosine-binding cassette (ABC) transporters at the cost of one ATP [54] (Figure 1). A type of enzyme-microbe synergy was recently reported for growing, metabolically active *C. thermocellum* cells that was attributed to surface phenomena involving adherent cellulolytic microorganisms rather than to the removal of hydrolysis products from the bulk fermentation broth [55]. In cell-free cellulase systems, another form of synergy exists between cellulases and extracellular β-glucosidases, which convert cellobiose and other cellodextrins to glucose. At high levels, cellobiose, one of the major products of cellulose hydrolysis, feedback inhibits cellulolytic activity [56], presumably to maintain a balance between cellulose degradation and the cell's ability to metabolize its catabolites. For *C. thermocellum*, cellobiose and longer cellodextrins are cleaved inside the cell either hydrolytically by β-glucosidases [57] or phosphorolytically by an intracellular cellobiose or cellodextrin phosphorylase [58-60] (Figure 1). The rate of the phosphorolytic cleavage reaction is about 20 times higher than the hydrolytic cleavage [61]. From a bioenergetic standpoint, phosphorolytic cleavage of imported β-glucan chains is preferable because the glucose-1-phosphate produced can be converted by phosphoglucomutase to glucose-

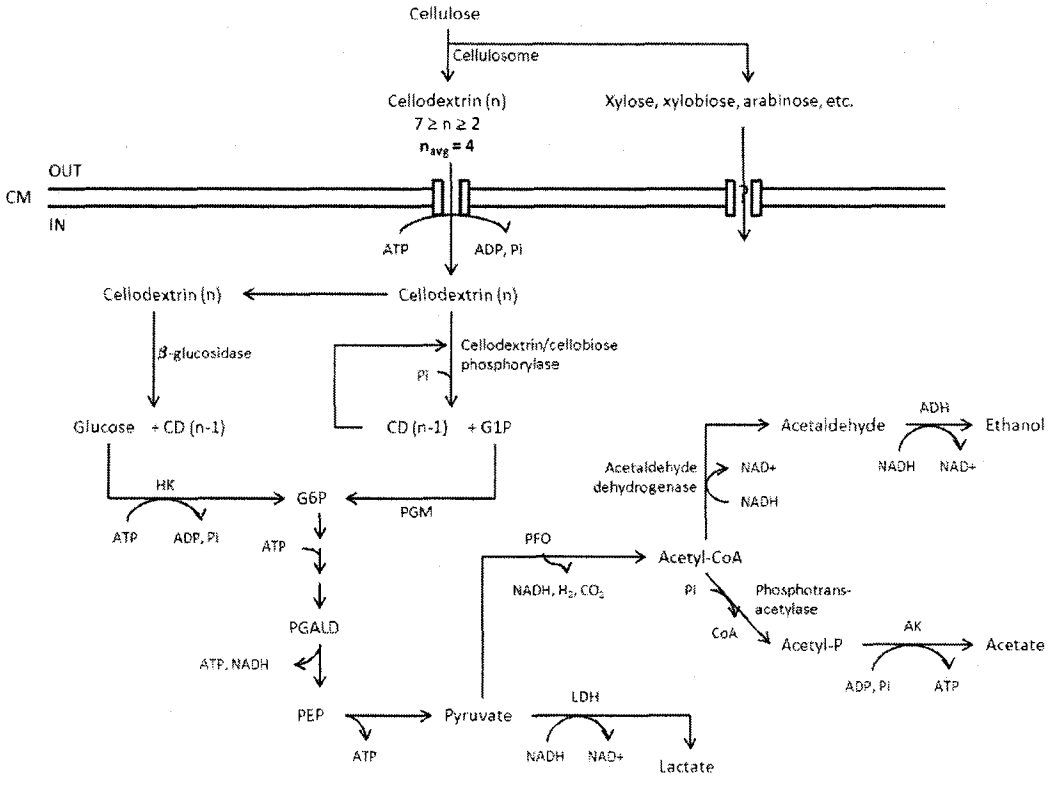

Figure 1. Cellulose utilization in *C. thermocellum*. Outside the cell, the action of the cellulosome converts cellulose to cellodextrins (CD) with degree of polymerization n, and also gives rise to products of hemicellulose degradation. Cellodextrins enter the cell via ABC transporters and, once inside the cell, they are cleaved either hydrolytically or phosphorolytically by β-glucosidase or cellodextrin phosphorylase, respectively. The former cleavage produces glucose which requires hexokinase (HK) to be converted to glucose-6-phosphate (G6P), while the former results in glucose-1-phosphate (G1P), which is converted to G6P by phosphoglucomutase (PGM). G6P enters glycolysis and pyruvate can be fermented to ethanol, acetate or lactate. CM, cell membrane; GPALD, 3-phosphoglyceraldehyde; PEP, phophoenolpyruvate; PFO, pyruvate ferredoxin oxidoreductase; ADH, alcohol dehydrogenase; AK, acetate kinase; LDH, lactate dehydrogenase.

6-phosphate for glycolysis and thus ATP production. Hydrolytic cleavage, on the other hand, produces only glucose, which costs an ATP in order to be converted to glucose-6-phosphate via hexokinase. It has been shown that the primary product taken up by the cell is a cellodextrin with on average 4 glucosyl moieties, not glucose or cellobiose [62]. With the energy savings on sugar transport that come with importing a cellodextrin with 4 degrees of polymerization and the benefits of the phosphorolytic cleavage of β-glucan bonds, this process was shown to be capable of supporting the cost of cellulase synthesis in anaerobes.

Cellulosome size is estimated at between 2×10^6 and 6×10^6 Da [63]. Assembled on the cell surface, polypeptides contain an N-terminal signal peptide that is cleaved off during secretion from the cell. Cellulosomes appear bound to the cell surface during log phase, become free in late exponential, and are mostly all free in stationary phase [63-65]. Cellulosomes have a requirement for Ca^{2+} and cellusomal activity is susceptible to oxidation due to the presence of sulfhydryl groups [66, 67].

2.2.2. Cellulosome structure

The structure of the *C. thermocellum* cellulosome consists in a central, noncatalytic, multimodular scaffolding protein bearing up to nine catalytic subunits (Figure 2) [68, 69]. The scaffolding protein is also referred to as scaffoldin or as (cellulosome integrating protein) CipA [69]. CipA has a predicted size of 196,800 [70], but it runs at higher than 200,000 by SDS-PAGE, likely because it is glycosylated [71]. The glycosylation may help protect the cellulosome from proteolytic cleavage in the extracellular environment [71, 72].

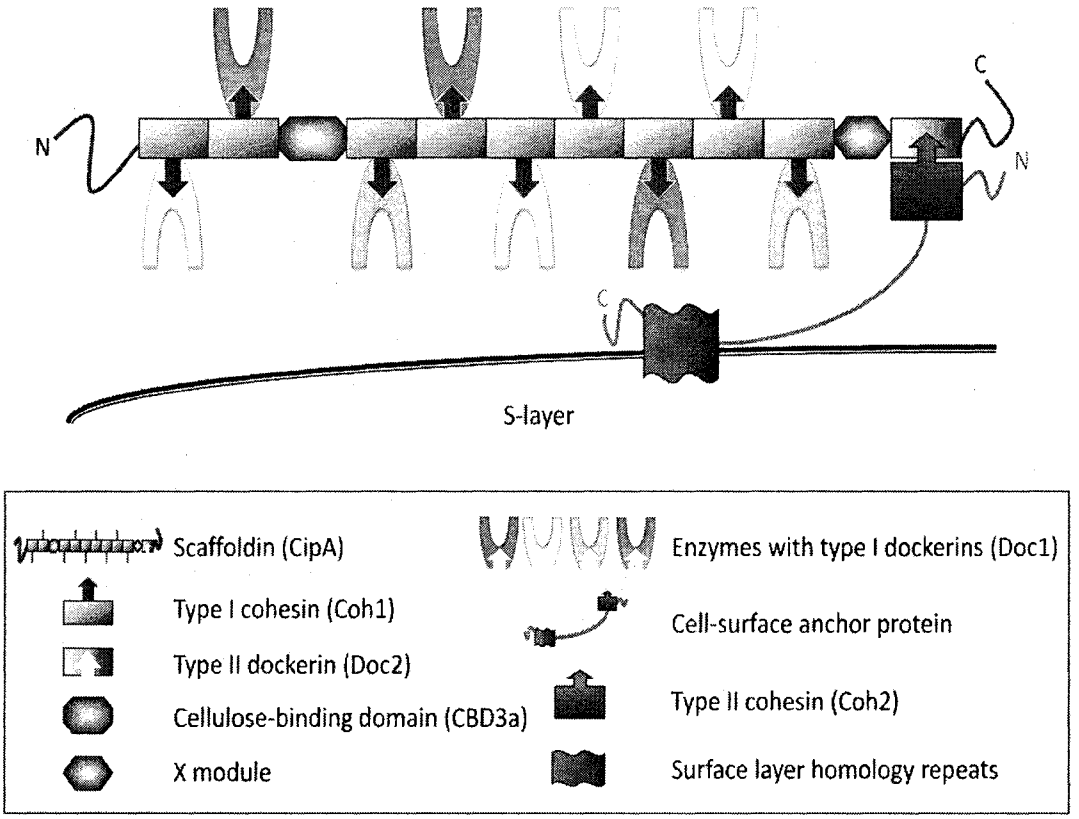

Figure 2. Structure of the *C. thermocellum* cellulosome complex. Scaffolding protein CipA binds 9 catalytic subunits via Coh1-Doc1 interactions. Doc2-Coh2 interactions mediate the binding of CipA to an anchor protein containing surface layer homology repeats that bind it noncovalently to the cell surface. In addition to 9 Coh1 domains, CipA contains a family IIIa cellulose-binding domain and a domain of unknown function.

The attachment of a given catalytic subunit to the cellulosome is mediated by the interaction of its type I dockerin (Doc1) domain with one of the nine highly conserved cohesin type I (Coh1) domains of CipA [73]. CipA contains, between its seventh and eighth Coh1 domains at the N-terminal, a type IIIa cellulose-binding domain (CBD3a), responsible for attachment of the complex and its enzymes to the surface of cellulose [70]. This mode of substrate targeting runs contrary to the cell-free fungal enzymes, which have their own CBDs for substrate binding. Some *C. thermocellum* cellulosomal enzymes do contain their own CBDs (CBD3b, CBD3c, CBD4, CBD30) but these do not bind cellulose as tightly as CBD3a [26]. While they may strengthen the binding to cellulose, their roles are more in facilitating the catalytic function of processive enzymes. The three-dimensional structure of a CBD3a revealed a 9-stranded β-sandwich with jelly roll topology and a Ca^{2+} binding site [74]. A comparison of the *C. thermocellum* CBD3a from CipA and CBDs from *Trichoderma reesei* showed that the former binds more sites on cellulose [75]. In addition to bringing the enzymes in contact with cellulose, the role of a CBD is believed to modify the surface of the substrate in order to promote hydrolysis [76, 77]. Some noncellulosomal enzymes in *C. thermocellum* have their own CBD3a [26].

CipA is bound to the cell-surface by virtue of the interaction of its C-terminal type II dockerin (Doc2) domain with the type II cohesin (Coh2) domain of one of three S-layer anchor proteins, SdbA, Orf2p, or OlpB [26]. SdbA has one Coh2 domain, Orf2p two Coh2 domains, and OlpB four Coh2 domains, presumably for binding one, two, and four CipA proteins, respectively. OlpA is a non-cellulosomal anchor protein that contains a Coh1 domain for tethering catalytic docking subunits directly to the cell surface [78].

These anchor proteins contain C-terminal surface-layer homology (SLH) repeats which integrate noncovalently with the S-layer (glycocalyx) just external to the peptidoglycan layer of the cell wall.

CipA also contains a module of unknown function, referred to as the X module, found between its first Coh1 and the Doc2 domain.

2.2.3. Cohesin and dockerin domains

The cohesin-dockerin interactions are crucial to complex formation in the cellulosome. The interactions are among the strongest noncovalent bonds found in nature. Binding assays using recombinantly expressed dockerin and cohesin polypeptides have been used to quantitate the thermodynamics of the interactions. Affinity constants on the order of between 10^9 and less than 10^{11} M^{-1} have been reported for both type I [79, 80] and type II interactions [81, 82], placing them in the high end of the range for typical protein-protein interactions [83]. The high affinity explains the remarkable stability of the quarternary structure of the cellulosome, which resists dissociation upon treatment with guanidine HCl, urea, nonionic detergents, and extremes in pH or ionic strength [84]. Treatment with SDS at temperatures above 70°C appears to consistently break the cellulosome into its component parts.

Cohesin-dockerin interactions are also highly type-specific in that Doc1 domains only recognize Coh1 domains, whereas Doc2 domains only recognize Coh2 domains [85]. Despite this type-specificity, there is no known specificity of particular Doc1 domains for any particular Coh1; thus, there is no known spatial order for binding of catalytic subunits along the CipA. Cohesin-dockerin interactions are also species-

specific. A Doc1 domain from *C. thermocellum* has been shown not to associate with a Coh1 domain from the CipC scaffoldin from *C. cellulolyticum* [86].

Structural studies in conjunction with mutagenesis approaches have uncovered the major aspects of cohesin-dockerin recognition and binding. The structures have been solved for Coh1, Doc1 and the Coh1-Doc1 complex (Figure 3). Like the CBD3a, the Coh1 fold is characterized by a nine-stranded β-sandwich with jellyroll topology [87, 88]. Repeats in the primary structure of the Doc1 sequence are manifested structurally as two Ca^{2+}-binding loop-helix motifs connected by a linker [89]. Proper folding of dockerin domains thus requires Ca^{2+}, hence the requirement for the divalent cation for cellulase activity. Both Ca^{2+}-binding segments of Doc1 are required for Coh1 recognition [80], which is mediated mainly by hydrophobic interactions between one of the faces of the Coh1 and α-helices 1 and 3 of the Doc1; Ser45 and Thr46 of the Doc1 dominate the hydrogen bonding network between it and the Coh1 [90].

The structures of Coh2 and Doc2 resemble their type I counterparts, however the Coh2-Doc2 complex, including the adjacent X module, showed that the latter participates in the interaction and may play a role in type I versus type II specificity [81].

2.2.4. Docking subunits: a variety of catalytic domains

Sequencing and annotation of the *C. thermocellum* ATCC 27405 genome led to the discovery of more than 60 open reading frames coding for products with putative Doc1 domains [91], that is, proteins that can potentially bind to CipA and contribute to

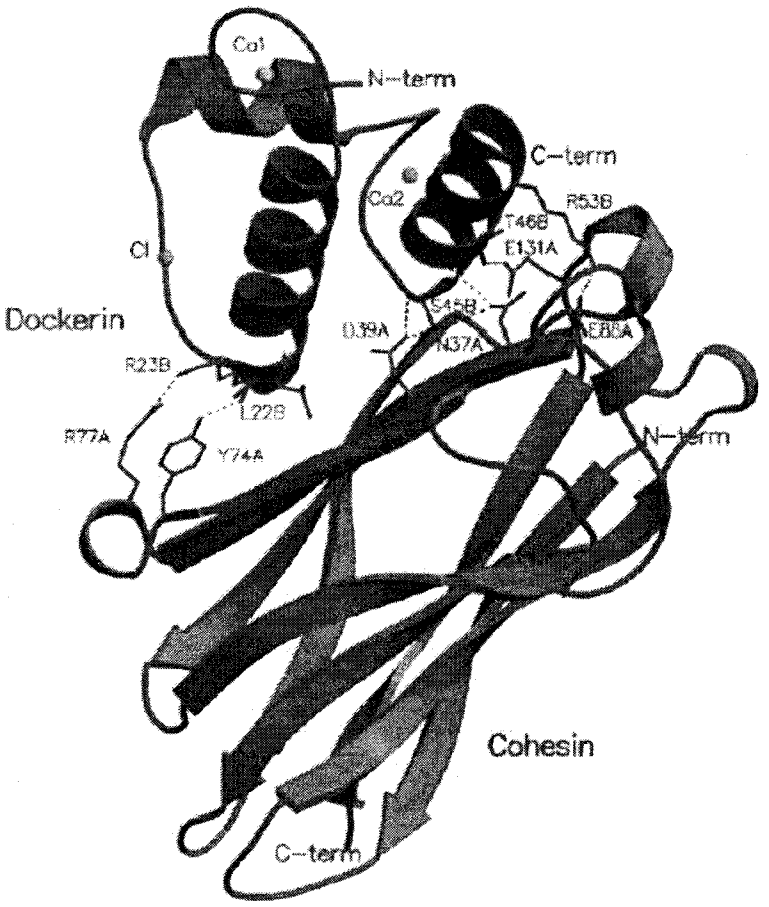

Figure 3. Structure of the type I cohesin-dockerin complex, reproduced from [90]. The complex is formed between the second Coh1 from CipA (red, lower right) and a Ca^{2+}-bound Doc1 (green, upper left). The residues involved in domain contacts are shown as stick models. The two Ca^{2+}-binding sites of the dockerin domain are shown as orange spheres.

cellulosomal activities. The predicted catalytic activity or function of about one-quarter of these genes is unknown. Considering the number of 'dockable' candidate open reading frames, relatively few, about one-third, of the products of these genes have been identified from the cellulosome complex itself. The participation in the cellulosome of the remaining putative gene products remains moot.

Twenty-seven docking component genes have been observed and/or cloned, expressed recombinantly and characterized: 4 that exhibit exoglucanase activity (CelS, CelK, CbhA, CelO), 12 with endoglucanase activity (CelA, CelB, CelD, CelE, CelF, CelG, CelH, CelJ, CelN, CelQ, CelR, CelT), 5 with xylanase activity (XynA, XynC, XynD, XynY, XynZ), one with chitinase activity (ChiA), one with mannanase activity (ManA), one with lichenase activity (LicB), one with xyloglucanase activity (XghA), and 2 that are nonenzymatic proteins (CseP, PinA). Both cellulolytic and hemicellulolytic glycoside hydrolases (GH) are classified into families according to the structural fold (predicted from primary structure) of the catalytic module [92], as are carbohydrate esterases (CE) [93]. Optimal conditions for these enzymes can range from pH 4.0-7.5 and temperatures of 55-78°C [94-100].

The major catalytic subunit of the cellulosome is the processive exo-acting cellobiohydrolase CelS [97, 101-103], which has a tunnel-shaped binding site [104] and is the only GH48 member in the *C. thermocellum* genome. Exoglucanases are the key enzymes in the degradation of crystalline cellulose [2], attacking cellulose chains from the reducing end, like CelS [104] or GH5 CelO [105], or the non-reducing end, like GH9 enzymes CelK and CbhA [106]. They work in concert with endoglucanases, which attack at random locations within a cellulose chain. Typically, exoglucanases exhibit higher

activity against crystalline forms of cellulose like Avicel or cotton, whereas endoglucanases prefer amorphous forms such as carboxymethyl cellulose (CMC). The major endoglucanase in *C. thermocellum* is CelA [12], the only member of GH8 in the *C. thermocellum* genome. The other endoglucanases have GH5 or GH9 folds. Many of the enzymes with GH9 folds also contain, directly C-terminal to the catalytic module, either a CBD3c (or an immunoglobulin-like domain), which has been shown to participate directly in the processivity of their catalytic function [107, 108].

The variety of hemicellulolytic activities is in keeping with the various types of hemicellulose that can exist in lignocellulosic materials. Six of the described docking components (CelE, CelH, CelJ, XynA, XynY, and XynZ) have more than one catalytic domain. Of note are the CE1 modules of XynY and XynZ. These have demonstrated feruloyl esterase activity, which would enable them to uncouple the cellulose-hemicellulose network from lignin [96].

Among the nonenzymatic docking proteins, PinA is a member of the serpin superfamily of serine protease inhibitors [109]. Presumably, it plays a role in defending the cellulosome against proteolytic cleavage in the extracellular environment. CseP bears sequence homology to spore-coat assembly protein CotH of *Bacillus subtilis*, which suggests it has a structural role in the cellulosome [110].

2.2.5. Regulation of cellulosomal enzymes

Cellulosome-related genes are regulated in a coordinated fashion to facilitate economic and efficient utilization of cellulosic materials [111]. Previous studies have shown that cellulolytic activity in *C. thermocellum* is regulated by either carbon source or

growth rate (or both), and that changes with respect to one or the other are reflected in overall cellulase production [112] and in the cellulosomal subunit profile [31, 84, 113, 114]. Expression of endoglucanases was observed to be controlled temporally, as *celA*, *celD* and *celF* transcripts were only observed at late exponential and early log phase during growth on cellobiose [115]. Catabolite repression by non-limiting concentrations of readily metabolized carbon sources has been the standing hypothesis for cellulase regulation in *C. thermocellum* for more than 20 years [30]. The catabolite repression scheme is supported by the presence of genes for Hpr, Hpr kinase, a CcpA-like LacI/GalR-family regulatory protein, and catabolite responsive element binding sequences in the *C. thermocellum* genome [112, 116]. While there is no evidence of a specific inducer being involved in cellulase synthesis, cellobiose does appear to be a repressor of genes responsible for activity against crystalline cellulose [117]; although overall endoglucanase activity is constitutive [30, 46, 118]. Higher cell-specific cellulase yields (mg per g of dry cell weight) are observed during growth on Avicel, and the decrease in cellulase yield has been correlated to increased extracellular cellobiose concentration [112]. The immediate availability of energy from cellobiose results in increased growth rate and leads to the repression of genes required to mine energy from crystalline cellulose. Lower growth rates and cellulose as substrate seem to promote cellulase production, as has been demonstrated for CelS, both at the protein [84] and the mRNA level [1, 119], as well as for the transcription of GH5 endoglucanases *celB* and *celG* and GH9 endoglucanase *celD* [120]. Transcription of scaffoldin gene *cipA* and cell-surface anchoring genes *olpB* and *orf2p* are likewise controlled by growth rate and/or carbon source, which is not the case for another cell-surface gene *sdbA* [119, 121].

Expression of xylanase *xynC* increases on cellobiose, both at the protein level [84] and the transcript level, although this increase does not appear to be growth-rate dependent [120]. Beyond these findings, hemicellulase regulation has not received much attention in *C. thermocellum*. In *Clostridium cellulovorans*, however, two major xylanases have been shown to be inducible by growth on xylan [122], and the mRNA levels of genes for xylanase *xynA* and pectinase *pelA* are also induced by growth in xylan- or pectin-containing media, respectively [123].

It has recently been shown that growth of *C. thermocellum* on laminaribiose induces genes for noncellulosomal β-1,3-glucanases CelC and LicA [124].

2.3. Mass spectrometry for quantitative proteomics

Low expression levels and overlapping and/or novel biochemical activity not detected by frequently used activity assays can account for the difference between the number of *C. thermocellum* cellulosomal proteins predicted and the number of those that have been biochemically characterized. Mass spectrometry (MS) has become an increasingly popular tool in the study of proteins due to its high sensitivity and mass accuracy, and its quantitative applications are being progressively refined [125]. The most wide-ranging *C. thermocellum* cellulosome study until now coupled a two-dimensional gel electrophoresis system with protein mass fingerprinting by matrix assisted laser desorption/ionization MS, giving rise to the simultaneous identification of 13 docking components from a cellulose-grown culture [91].

2.3.1. Peptide sequencing by MS/MS

Shotgun MS approaches have been developed for the study of entire proteomes or subproteomes [126-129]. A mixture of proteins is digested to peptides with trypsin. The digested protein in solution phase can be resolved into its constituent peptides, according to their relative hydrophobicities, by LC placed in-line with a tandem mass spectrometer. Electrospray ionization (ESI), a soft ionization technique used so that large peptides do not break apart prior to MS/MS fragmentation, converts the eluting peptide ions from solution to gas phase by pushing the liquid through a very narrow capillary to which a charge is applied. In an ion trap MS such as the Thermo LTQ (Figure 4), the ions produced by ESI are focused into the ion trap by an electrostatic lensing system (ion optics). An ion gate system pulses open and closed to allow ions into the trap and confine them by creating a potential well [127]. Ions in the trap can be released selectively, leaving behind only the precursor ion of interest to be dissociated by collisional activation with helium as damping gas (collision induced dissociation or CID), which converts kinetic energy to vibrational energy resulting in fragmentation [127]. Fragment ion products are ejected from the trap and detected using an electron multiplier. The MS/MS spectrum of fragment-ion peaks generated reflects the amino acid sequence of the precursor peptide. The peptide sequence is established from the mass differences between the peaks, using b- and y-type ions, which extend from the amino and the carboxy termini, respectively (Figure 5). This information is recorded as a list of the peptide fragment masses and their intensities (stored as a DTA file by Thermo Electron software). This list is then matched to a theoretical peptide fragment spectrum in a sequence database, which contains the masses and intensities of peptide fragments from a

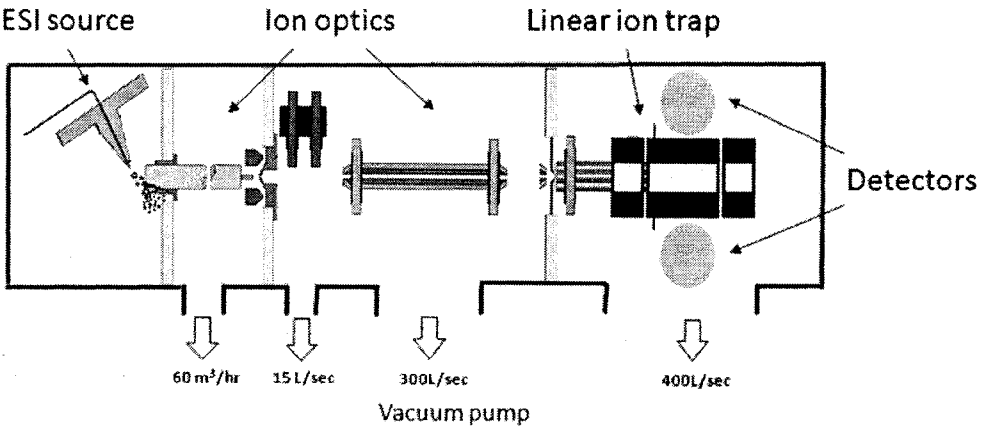

Figure 4. Schematic of the LTQ linear ion trap mass spectrometer.

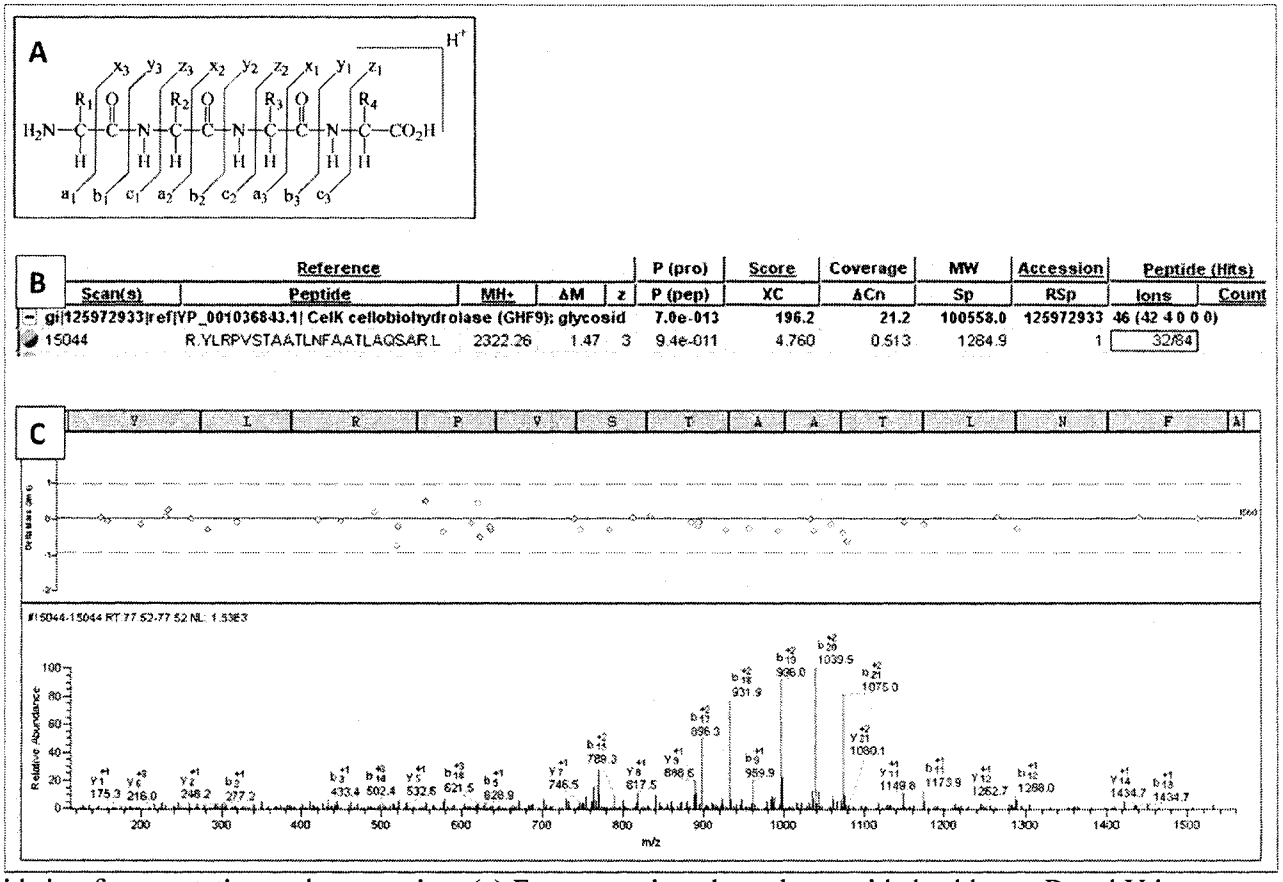

Figure 5. Peptide ion fragmentation and sequencing. (a) Fragmentation along the peptide backbone. B and Y ions are most commonly used for peptide sequencing. B or Y ions result from the proton migrating to the amino or the carboxy terminus, respectively. (b) Peptide/protein hit information in BioWorks 3.3 browser for the sequenced peptide YLRPVSTAATLNFAATLAQSAR, detected in triply charged form at m/z 775.25. With 32 of 84 possible b- and y-type ions detected, this sequence was assigned a XC score of 4.760. (c) The corresponding MS/MS scan data for the fragmentation of YLRPVSTAATLNFAATLAQSAR, with b and y ions shown.

collection of proteins or translated open reading frames digested with trypsin *in silico* [130]. In such a way, multiple peptides can be detected and correlated to a protein. The correlations can be calculated using the SEQUEST algorithm, which assigns a cross correlation score (XC) that takes into account the percentage of fragment ions detected for a peptide sequence, as well as the number of peptides identified per protein [131]. The more peptides are sequenced, the higher the confidence in the protein correlation. Thermo Electron's BioWorks software also calculates a P_{pro} value which represents the likelihood that the sequence information should correlate to another protein in the sequence database. Given trypsin specificity, intact peptide mass, and a partial amino acid sequence, the protein correlation can be very strong even with a single peptide, in contrast to a peptide mass fingerprinting experiment.

2.3.2. Relative quantitation using internal standards

Relative quantitation of peptides/proteins can be done in nanoLC-ESI-MS experiments with the use of internal standards. The absolute signal intensity of a peptide ion measured by MS does not necessarily reflect the abundance of that peptide in a mixture with other peptides. Two different peptide sequences present in equal abundance can give signals of unequal intensity in a single MS run. This is due to differences in ionization efficiencies between peptide ions and to background and ion suppression effects. The use of an internal standard accounts for these effects, and also controls for losses that occur during sample preparation (if added prior to extraction) and LC injection. The best internal standard is an isotopically labelled version of the peptide to be quantified. An isotopically labelled internal standard will have a similar extraction

recovery, chromatographic retention time, and ionization response in ESI-MS as its unlabelled analog. MS is particularly well suited for the use of stable-isotope labelled internal standards because of its ability to measure masses at high accuracy; however, the labelled peptide should provide a difference of at least 3 Da for adequate separation from the naturally occurring isotopic distribution around the peptide ion being measured [132]. Several quantitative proteomics technologies exist that involve incorporation of stable isotope tags either *in vivo* or *in vitro*. *In vitro* labelling can be done after the protein is digested. ITRAQ (isobaric tags for relative and absolute quantitation) technology involves chemically modifying the amino termini of peptides with stable isotope-labelled reagents [133]. Labels can also be added during protein digestion with trypsin. During proteolysis, trypsin incorporates an oxygen atom from the surrounding water. Performing the digestion in ^{18}O water incorporates a 2-Da difference per peptide created [134]. Then again, protein can be tagged prior to digestion. ICAT (isotope-coded affinity tag) technology involves labelling the cysteine residues in proteins with a stable isotope label [135]. Alternatively, labels can be incorporated *in vivo*. SILAC (stable isotope labelling with amino acids in cell culture) involves growing cells in medium containing stable-isotope labelled amino acids. Instead of the amino acids being labelled, it could be some other reagent in the growth medium like a ^{13}C carbon source or ^{15}N nitrogen source (Figure 6) [136, 137]. When the labelled analog of the carbon or nitrogen source is supplied to cells in culture, it gets incorporated into all newly synthesized proteins. After a number of cell divisions, all instances of the original will be replaced by the analog. Since there is hardly any chemical difference between the labelled and the natural isotopes, the cells behave similarly.

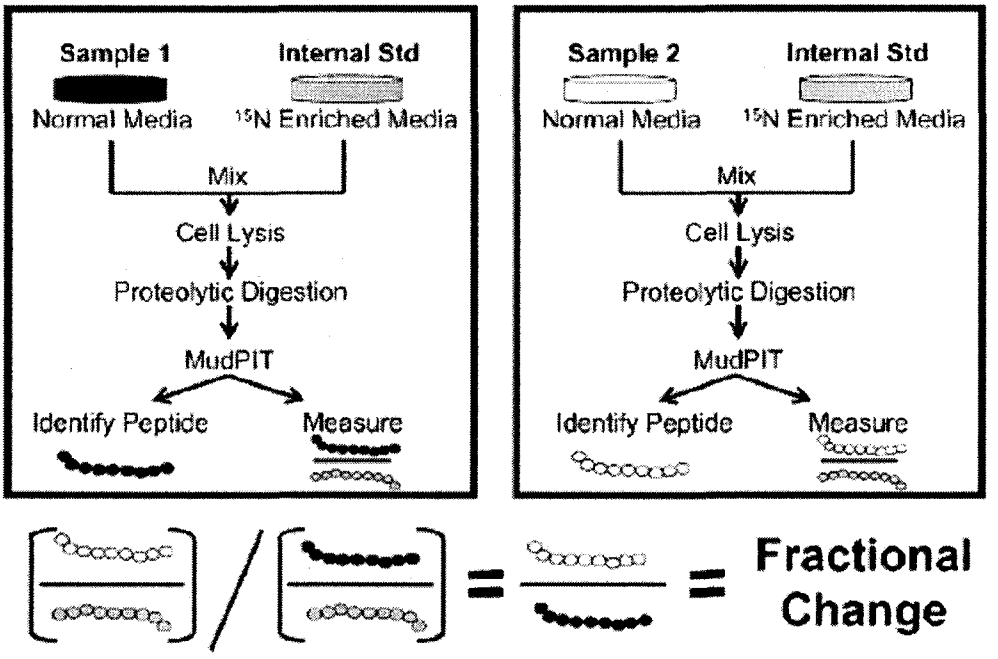

Figure 6. General scheme for quantitative proteomics using metabolic labelling, reproduced from [137]. Cells grown in media containing ^{15}N are mixed with cells from different conditions prior to protein extraction and digestion. Measurement of the proteins from each sample is made using their respective ^{15}N-labelled protein as internal standards. Changes in protein level are expressed relative to another sample to minimize systematic errors.

In all of these strategies, one of two (or several) samples being compared is labelled, and then mixed with unlabelled sample (or sample tagged with a different label). In extracting quantitative data, mass spectra are acquired, resulting in isotope clusters for each pair of labelled and unlabelled peptides. As the peptides co-elute from the column, their signals are sampled several times, tracing out individual ion-current curves (Figure 7). The area under each curve is an extracted ion chromatogram (XIC) and is proportional to the peptide's abundance [125]. Differences in abundance can be determined by comparing the area under each peak in a ratio. A complex mixture analysis can yield thousands of peptide XICs which can be correlated to proteins and used for quantitation of their relative abundance. A computer program called RelEx has been developed for the calculation of such peptide ion-current ratios using a least-squares regression (Figure 7) [137].

2.3.3. Quantitation by peptide counting methods

An altogether different approach, not relying on internal standards for the MS-quantitation of proteins, involves peptide counting. Such methods correlate the number of peptides detected per protein to the abundance of that protein in a mixture with other digested proteins. The need to normalize this number somehow becomes clear when it is considered that a large and a small protein present in a mixture in equal concentration do not yield the same number of peptides upon proteolysis. Normalizing to the number of theoretical peptides, peptides that can be detected within the LC run and the mass range of the MS, yields a rough proportionality to protein abundance, as per the emPAI (exponentially modified protein abundance index) method [138]. The concept of theoretically LC-MS-observable peptides has evolved into the notion of a proteotypic

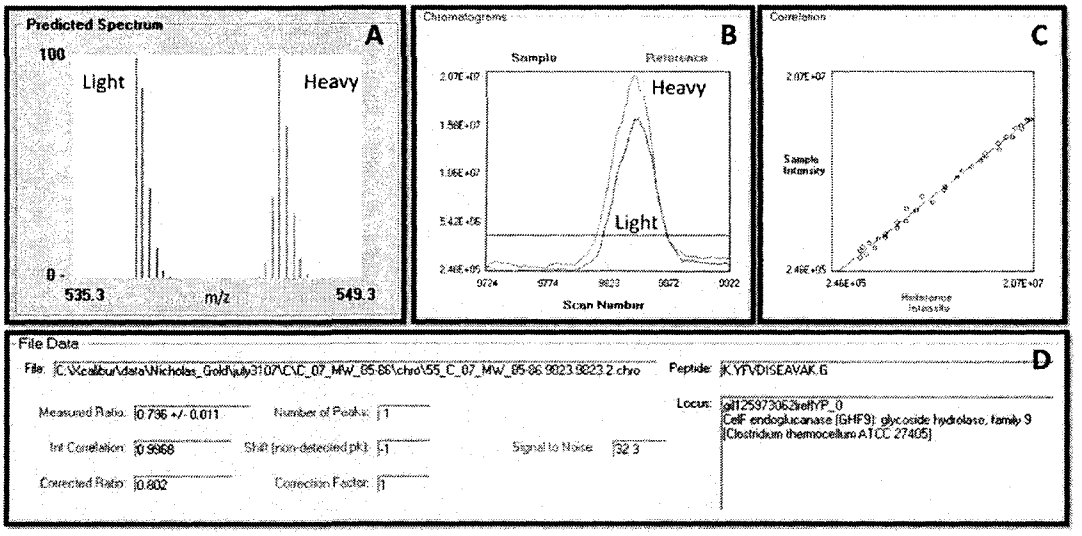

Figure 7. Graphical user interface for RelEx software upon analysis of peptide to labelled peptide ratios [137]. (a) Mass spectra are acquired, resulting in isotope clusters for each peptide, the naturally occurring isotope distribution (left) and the labelled peptide distribution (right). (b) As the peptides co-elute from the column, their signals are sampled several times, tracing out overlapping extracted ion chromatograms (XICs) for each. The area under each curve is proportional to that peptide's abundance. (c) RelEx determines a correlation factor as a measure of the overlap of the XICs. (d) Differences in abundance are determined by calculating the ratio of the areas under each curve.

peptide, which is an experimentally observable peptide that uniquely indentifies a specific protein [139]. Proteotypic peptides are being used to normalize emerging peptide counting quantitation methods [140-142].

3. METHODS

The general scheme for the comparison of cellulosomal subunit profiles from *C. thermocellum* cells grown using different carbon sources is depicted in Figure 8. In summary, *C. thermocellum* was grown on different substrates in liquid (batch) culture. Each sample culture was mixed with a reference culture in which all proteins were labelled metabolically with ^{15}N. Cellulosomes were isolated from each mixture, separated by SDS-PAGE, and then digested with trypsin for peptide sequencing by nanoLC-ESI-MS/MS. Cellulosomal proteins were identified by using the SEQUEST algorithm to match (unlabelled) peptide sequence information to the *C. thermocellum* sequence database. The unique (unlabelled) peptides observed were counted and used in the calculation of relative protein abundances per sample by the emPAI method. Labelled peptides acted as internal standards for the determination of relative differences in protein abundance between two samples, using RelEx software.

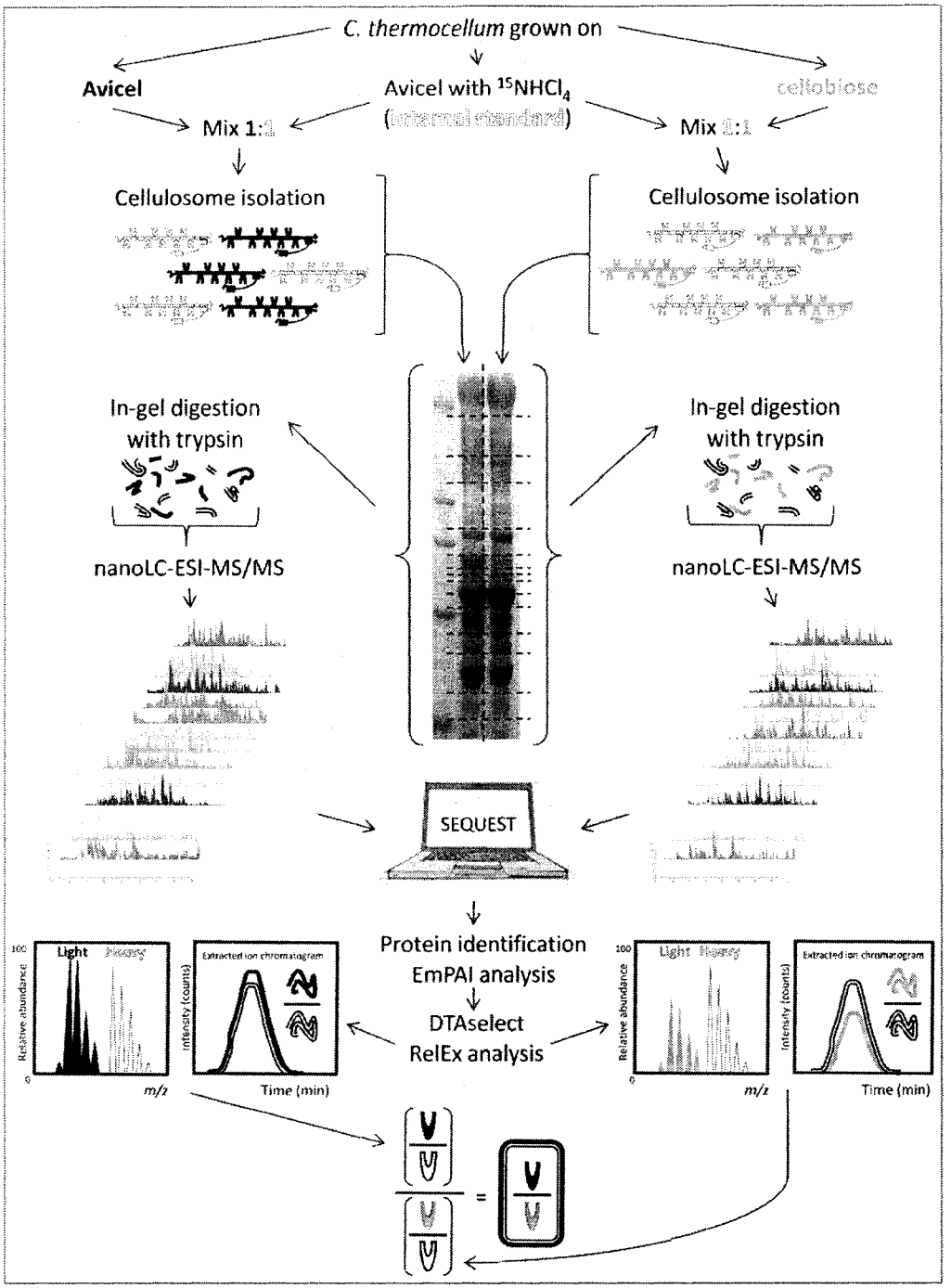

Figure 8. General scheme for the comparison of cellulosomes from Avicel- and cellobiose-grown *C. thermocellum* cells. Cells from each sample culture were mixed with an internal standard culture grown in medium enriched with ^{15}N. Cellulosomes were isolated from each mixture, separated by SDS-PAGE, digested proteolytically with trypsin, and then analyzed by nanoLC-ESI-MS. Proteins were identified by matching MS/MS spectra to the *C. thermocellum* sequence database using SEQUEST. Protein abundances were evaluated by the emPAI method and by RelEx analysis.

3.1. Media preparation for growing *C. thermocellum*

The liquid media were based on ATCC medium 1191, but without sodium sulfide. They were prepared from the mixture of three separate solutions: a buffer solution, a vitamin solution and a mineral solution. The buffer solution was prepared by combining the ingredients listed in Table 1. The dry ingredients were dissolved completely before adding sodium hydroxide. The solution was then transferred to anaerobic culture bottles (Bellco Glass) in volumes of 95 mL. When Avicel PH101 (Fluka-Biochemika) was used as carbon source, 200 mg was added to each bottle (final concentration of 0.2%, wt/vol). Xylan from birch wood (100 mg; Sigma Aldrich), pectin from citrus peel (50 mg; Sigma Aldrich), and locust bean gum (50 mg from *Ceratonia siliqua* seeds; Sigma Aldrich), when used, were also added at this point. Solutions were sparged in the anaerobic bottles with nitrogen gas for 3-5 min, and then quickly stoppered with a rubber septum which was then sealed with an aluminum cap. Sealed media bottles were then sterilized by autoclaving on liquid cycle for 15 min.

The vitamin solution was prepared by combining the ingredients in Table 2. Only small amounts of the vitamin solution are required, so unused vitamin solution was divided into 50-mL aliquots and frozen at -20°C. In preparing the mineral solution from the ingredients in Table 3, the nitrilotriacetic acid was first suspended in 500 mL of water, and then titrated to pH 6.5 with 2 N KOH to dissolve. Unused mineral solution was filter-sterilized into an autoclaved bottle and left at room temperature for later use. A mixture of the vitamin and mineral solutions was prepared by combining 1 mL of the former with 10 mL of the latter, and diluting up to 100 mL. When cellobiose (Sigma-Aldrich) was used as carbon source, 4 g was dissolved into this vitamin-mineral solution,

Table 1. Buffer solution for ATCC 1191 liquid growth medium

	Per L of water	
	Full medium	Minimal medium
KH_2PO_4	1.58 g	1.58 g
$Na_2HPO_4 \cdot 12H_2O$	4.42 g	4.42 g
$^{14}NH_4Cl$	0.53 g	
$^{15}NH_4Cl$ (99%)		0.53 g
$MgCl_2 \cdot 6H_2O$	0.19 g	0.19 g
L-cysteine HCl	0.53 g	0.53 g
Yeast extract	2.11 g	
Resazurin (0.1% wt/vol)	1.05 mL	1.05 mL
NaOH (10N)	842 μL	842 μL

Table 2. Vitamin solution for ATCC 1191 liquid growth medium

	Per L of water	
	Full medium	Minimal medium
Biotin	40 mg	40 mg
p-Aminobenzoic acid	100 mg	100 mg
Folic acid	40 mg	40 mg
Pantothenic acid calcium salt	100 mg	100 mg
Nicotinic acid	100 mg	100 mg
Vitamin B12	2 mg	2 mg
Thiamine HCl	10 mg	10 mg
Pyridoxine HCl	200 mg	
Pyridoxal HCl		200 mg
Thioctic acid	100 mg	100 mg
Riboflavin	10 mg	10 mg

Table 3. Mineral solution for ATCC 1191 liquid growth medium

	Per L of water
Nitrilotriacetic acid	1.5 g
$MgSO_4 \cdot 7H2O$	3 g
$MnSO_4 \cdot H_2O$	500 mg
NaCl	1 g
$FeSO_4 \cdot 7H_2O$	100 mg
$Co(NO_3)_2 \cdot 6H_2O$	100 mg
$CaCl_2$ (anhydrous)	100 mg
$ZnSO_4 \cdot 7H_2O$	100 mg
$CuSO_4 \cdot 5H_2O$	10 mg
$AlK(SO_4)_2$ (anhydrous)	10 mg
Boric acid	10 mg
$Na_2MoO_4 \cdot 2H_2O$	10 mg
Na_2SeO_3 (anhydrous)	1 mg

such that the final concentration was 4% (wt/vol). The vitamin-mineral mixture, with or without cellobiose, was sparged with nitrogen for several minutes. While sparging, the solution was drawn up through a stainless steel cannula into a 60-mL syringe. The cannula was replaced with a 0.2 μm filter to which was fixed a syringe needle. The solution was then filter-sterilized into a clean, empty anaerobic culture bottle, which had been previously autoclaved, flushed with nitrogen gas, and stoppered as before. Five mL of this solution were added to the buffer solution, for a total volume of 100 mL per bottle (final cellobiose concentration of 0.2%, wt/vol).

3.2. Growth conditions and metabolic labelling

For comparison of growth on cellulose versus growth on cellobiose, *C. thermocellum* strain ATCC 27405 was grown anaerobically at 58°C in 100-mL batch cultures in full ATCC medium 1191, containing 0.2% (wt/vol) of either Avicel (the model substrate for crystalline cellulose) or cellobiose. An Avicel-grown reference culture was prepared similarly in minimal ATCC medium 1191, in which 99% ^{15}N-enriched NH_4Cl (Cambridge Isotope Laboratories, Andover, MA) was substituted for the nitrogen source and pyridoxine HCl was replaced with pyridoxal HCl. A 5% (vol/vol) inoculum of unlabelled Avicel-grown cells was passed three times into $^{15}NH_4Cl$-containing medium, before inoculation of the final reference batch, which was consequently enriched with ^{15}N to an estimated 98.9%. All cultures were harvested for protein isolation in late stationary phase (70 h), at which point each test culture was mixed 1:1 (vol/vol) with the reference culture.

For comparison of growth on either cellulose or cellobiose with and without hemicelluloses, Avicel- and cellobiose-grown cultures of *C. thermocellum* strain ATCC 27405 in exponential phase, grown as above, were used to make 3% (vol/vol) inoculae into Avicel and cellobiose media, respectively, containing 0.1% (wt/vol) xylan (β-1,4-linked xylose), 0.05% (wt/vol) pectin (α-1,4-linked galacturonic acid), and 0.05% (wt/vol) locust bean gum (β-1,4-linked mannose with occasional galactose branch points). Xylan (X), pectin (P) and locust bean gum (M) will often be referred to collectively in the text as XPM. A reference culture was grown and enriched as above except with XPM added to the Avicel-containing minimal medium, which resulted in a growth lag. The reference culture was therefore inoculated 48 h prior to the test cultures, which were harvested at once in late stationary phase (70 h). All test cultures were mixed 1:1 (vol/vol) with the reference culture.

3.3. Protein fractionation

The steps in the isolation of various *C. thermocellum* protein fractions are shown in Figure 9. A 1-L culture grown on cellobiose to stationary phase, as above, but not mixed with a reference culture, was centrifuged at 10,000 × *g* for 10 min. The supernatant was divided into two portions. To one portion was added 4 volumes of cold acetone, and the mixture was left at 4°C for 30 min to precipitate the total extracellular protein. The mixture was centrifuged at 17,000 × *g* for 20 min to pellet the protein, which was suspended in 50 mM Tris-HCl, pH 7.4. Cellulose-binding extracellular protein (the cellulosome fraction) was removed from the other portion of the supernatant (as

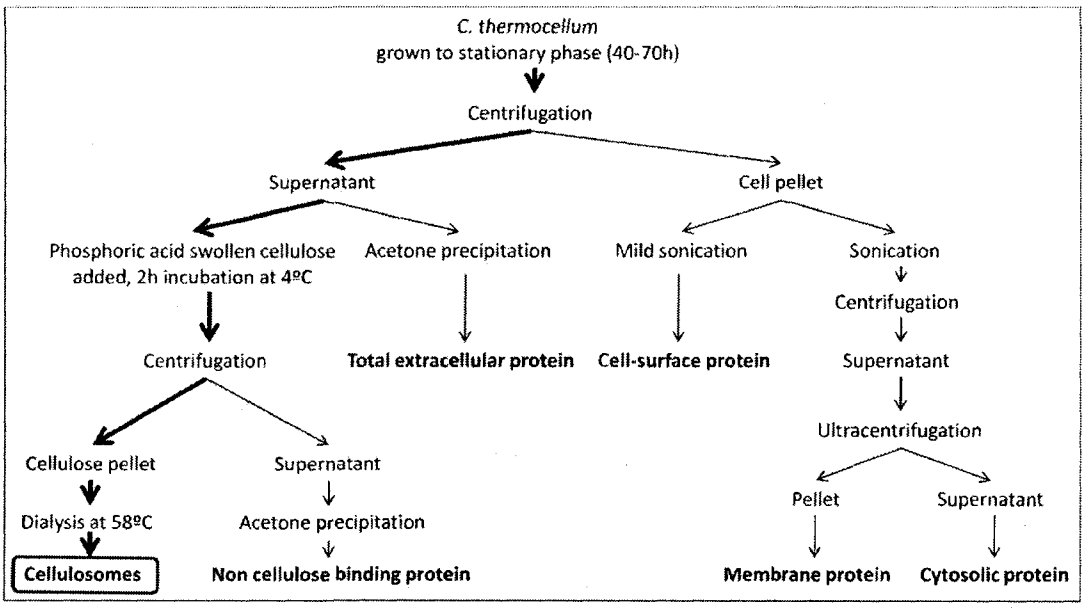

Figure 9. General protein fractionation scheme. The left-most path follows the isolation of cellulosomes from *C. thermocellum* grown to stationary phase in liquid culture.

described in section 3.5), and the non cellulose-binding extracellular protein remaining in the supernatant was acetone-precipitated and recovered as before.

The cell pellet was divided into three portions that were treated by different methods in attempts to effect the release of proteins from the cell surface. Cells were suspended in 50 mM Tris-HCl, pH 8.0, and sonicated at 0.3 cycle/s with pauses of 0.5 s and a power setting of 0.5 (maximum strength 550 W) [143]. Other cells were suspended in a 50 mM Tris-HCl sucrose buffer, pH 8.0, containing 1 mg/mL lysozyme, 250 μg/mL RNAse A, and 2 mM phenylmethylsulfonyl fluoride, and incubated at 37°C for 1 h [144]. Finally, cells were suspended in 50 mM Tris-HCl, pH 8.0, with 8M urea, and incubated at room temperature for 30 min [144]. After all three treatments, cells werepelleted again by centrifugation, and the supernatants were concentrated by acetone precipitation as above.

The total extracellular protein, cellulose-binding extracellular protein, non-cellulose binding extracellular protein, and three surface-layer protein fractions were resolved by SDS-PAGE (6%) and stained with Coomassie Blue.

3.4. Preparation of phosphoric acid swollen cellulose

Phosphoric acid swollen cellulose (PASC) was prepared as per Walseth [145] with some modifications. In a mortar kept on ice, 5 g of Avicel was gradually added to 100 mL of phosphoric acid (85%), with stirring using a pestle to avoid lumps. Once all the Avicel was added, the mixture was stored at 4°C for 30 min to allow swelling. The mixture was then washed several times with cold water and then Tris-HCl buffer (50 mM, pH 6.8), with centrifugation steps in between, until the mixture had reached a stable pH of 6.8. Finally, the mixture was homogenized in a blender to remove lumps.

3.5. Isolation of cellulosomes by affinity digestion

Supernatants were collected by centrifuging sample cultures or sample-reference culture mixtures at 10,000 × *g* for 10 min. The pH of the supernatants was adjusted to 6.8 with 10 N NaOH. To 900 mL of each were then added approximately 14 mg of PASC. The supernatants containing PASC were incubated at 4°C with stirring for a minimum of 2 h. PASC pellets were then collected by centrifugation for 20 min at 17,000 × *g*. Pellets were washed once with cold water, re-centrifuged and then suspended in 5-7 mL of dialysis buffer consisting of 5 mM dithiothreitol (DTT), 0.1 g/L L-cysteine HCl, 2 mM EDTA, 12 mM $CaCl_2$, and 50 mM Tris-HCl, pH 6.8. The suspensions were then transferred to Pierce Slide-A-Lyzer dialysis cassettes (MWCO 10,000), which were each placed in 1 L of water held at 58°C while stirring on a hot/stir plate. This procedure, termed 'affinity digestion' and first developed by Morag et al. [146], is illustrated in Figure 10. Dialysis is necessary because, as the digestion progresses, cellulases active at high temperature degrade the PASC, releasing cellobiose which inhibits cellulolytic activity [56]. The digestion-dialysis should not be allowed to carry on for too long, for otherwise, should the cellulases finish breaking down the PASC, they will attack the dialysis membrane, which itself is made of nitrocellulose; and this would result in total loss of sample. After a 5-h digestion and dialysis period at 58°C, the contents of the cassettes were removed and precipitated with four volumes of cold acetone. The precipitates were collected by centrifugation, dried down in a vacuum centrifuge, and suspended in 50 mM Tris-HCl, pH 7.4, each to final concentrations of approximately 10 mg/mL, as verified by Bradford protein assay.

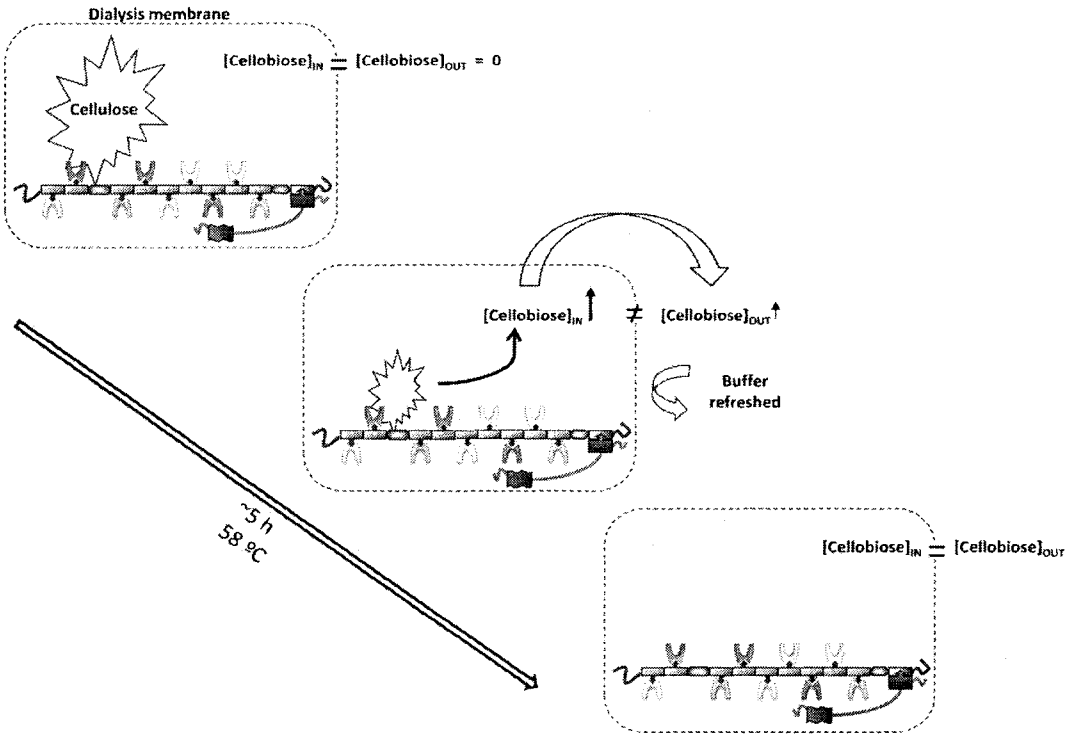

Figure 10. Affinity digestion method for cellulosome isolation from culture supernatant. Cellulose-bound protein is suspended in buffer containing Ca^{2+} and a reductant, and placed in a dialysis cassette. Over the course of a 5-h digestion-dialysis period against water at 58°C, cellulose is converted to cellobiose which is removed by osmosis. Purified cellulose-binding protein remains.

3.6. Analysis of gel-separated cellulosomes by nanoLC-ESI-MS

Purified cellulosome mixtures were separated by 6% SDS-PAGE and stained with Coomassie Blue. Sample lanes from the gel were excised and divided into fifteen gel bands, with each band containing on average roughly 11 μg of protein. The protein in each gel band was subsequently reduced and alkylated (to prevent reduced cysteine thiols from forming oxidized disulfide bonds), and digested with trypsin TPCK (Sigma-Aldrich), as described previously [147]. (See Appendix C for in-gel digestion protocol.) The resulting peptide mixtures were removed from the gel pieces using excess extraction buffer, dried, and then made up in equal volumes of 8% (vol/vol) acetonitrile (ACN) in 0.1% (vol/vol) formic acid (FA). (Alternatively, proteins can be digested with trypsin in-solution; see Appendix D for protocol.) Peptide samples were injected quantitatively for separation on a PicoFrit BioBasic C18 nanocolumn (New Objective; 10 cm length × 75 μm inner diameter, 5 μm particle size, 300 Å pore size) with a 60-min solvent gradient, ranging from 3% to 50% ACN in 0.1% FA, at a flow rate of 1 $\mu L \cdot min^{-1}$. Before flowing to the column, sample was cleaned of impurities using a C18 peptide trap. Under these conditions, most peptides eluted in about 30 s or 500 nL. Detection and sequencing of peptide ions was accomplished by an LTQ linear ion-trap MS (Thermo Electron, San Jose, CA USA), equipped with an ESI nanosource and operating in positive mode with a voltage of 1.4 kV applied at a liquid junction just upstream of the column. Initial full MS survey scan (~10 ms) was performed for the m/z range of 400-2000, followed by several data dependent scans (~33 ms each). The seven most abundant ions from the survey scan were subjected to MS/MS for sequencing using pulsed-Q dissociation for ion fragmentation. A triggering threshold of three times the noise level (S/N) was applied for

MS/MS events. Peptide ions that triggered an MS/MS more than once within a 30-s window were placed on an exclusion list for three minutes to improve the possibility of less abundant ions being detected.

For comparison of cellulosomes from cells grown on cellulose or cellobiose with and without XPM, rather than a C18-packed nanocolumn, a BioBasic C18 column with 180-μm internal diameter was used in conjunction with an unpacked nanocolumn for respective peptide separation and ESI. The solvent gradient ran from 8-40% ACN in 0.1% FA in 90 min.

3.7. Database screening and success criteria

Using SEQUEST from BioWorks 3.3 (Thermo Electron), peptide sequence results were searched against the 2007/02/16 release of the *C. thermocellum* genome available at the National Center for Biotechnology (NCBI) website courtesy of the U.S. Department of Energy (DOE), JGI (http://www.ncbi.nlm.nih.gov, Refseq accession number NC_009012). The database was digested *in silico* with trypsin, generating peptides within the mass range 400-3500 Da. Furthermore, the database was indexed for a maximum of 3 of the following post-translational modifications (PTMs) per peptide: carboxymethylation of cysteine residues (monoisotopic δ mass of 58.0050), oxidation of methionine residues (to sulfoxide, δ mass of 15.99490), N-terminal acetylation and acetylation of lysine residues (δ mass of 42.01060). A peptide tolerance of ± 2 atomic mass units was implemented. Charge state analysis was performed during DTA file filtering, and a series of high-stringency filters were applied to the search results. Singly, doubly and triply charged peptide ions required SEQUEST cross correlation (XC) scores

of at least 1.8, 2.5, and 3.5, respectively. Peptide and protein hits also needed probability scores, as calculated by BioWorks, of less than 10^{-3}. Moreover, only proteins identified on the basis of two or more unique peptides were considered in the final analysis. The SignalP 3.0 server (http://www.cbs.dtu.dk/services/SignalP/) was used to verify that proteins contained an N-terminal peptide signalling secretion from the cell [148].

3.8. RelEx analysis

DTA files were filtered separately using DTASelect [149], which assembles the peptides into proteins using the same XC-score stringency factors as above. The filtered DTA files were then analyzed by RelEx [137], which generates extracted ion chromatograms of peptide isotope pairs, and uses the areas under each curve to calculate a peptide signal ratio of sample to isotope-labelled reference. (See Appendix F for DTASelect-RelEx procedure.) An extracted ion chromatogram pair was rejected if the S/N was below three or if the correlation factor, the measure of the overlap of the curves, was below 0.9. Protein ratios were calculated as averages of the ratios of the peptides matched to them. The ratio of each unlabelled Avicel-grown protein over ^{15}N-labelled Avicel-grown protein was divided by the ratio of the corresponding unlabelled cellobiose-grown protein over ^{15}N-labelled Avicel-grown protein. The quotient of the ratios is the ratio of unlabelled Avicel-grown protein over cellobiose-grown protein. In such a way, this strategy corrects for any systematic errors introduced during sample preparation [137]. All ratios were normalized to that obtained for the comparison of CipA. Given that the time required for a single measurement places practical limits on the number of replicate values of individual samples that can be performed in determining

error and the significance of observed changes in protein abundances, assessing variance by multiple peptides per protein in a single run is an acceptable alternative that approximates the same result [132]. Standard error (SE) in the normalized ratio of ratios was calculated using the simple rules of error propagation for quotients. Since

Eq. 1

$$R_A = \frac{\left(\frac{r_{A,Avicel}}{r_{A,cellobiose}}\right)}{\left(\frac{r_{CipA,Avicel}}{r_{CipA,cellobiose}}\right)}$$

where R_A is the overall ratio of sample to reference ratios r, normalized to CipA, for a given protein A, then the overall standard error in R_A is

Eq. 2

$$SE_A = R_A \times \sqrt{\left(\frac{SD_{A,Avicel}}{r_{A,Avicel}}\right)^2 + \left(\frac{SD_{A,cellobiose}}{r_{A,cellobiose}}\right)^2 + \left(\frac{SD_{CipA,Avicel}}{r_{CipA,Avicel}}\right)^2 + \left(\frac{SD_{CipA,cellobiose}}{r_{A,cellobiose}}\right)^2}$$

where uncertainties (standard deviations SD) in A on Avicel, A on cellobiose, CipA on Avicel, and CipA on cellobiose are random and independent.

The two-tailed Student's t-test was used to determine the probability that the ratios calculated for growth on Avicel and for growth on cellobiose corresponded to two distinct populations between which real differences could be observed. The t-distribution value was calculated as

Eq. 3

$$\left| \frac{r_{A,Avicel} - r_{A,cellobiose}}{\sqrt{\left(\frac{1}{N_{A,Avicel}} + \frac{1}{N_{A,cellobiose}}\right)\frac{\left[(N_{A,Avicel}-1)\times SD^2_{A,Avicel} + (N_{A,cellobiose}-1)\times SD^2_{A,cellobiose}\right]}{df}}} \right|$$

where N is the number of peptides used for the calculation of the ratio r, and df is the degrees of freedom, which is

Eq.4

$$df = N_{A,Avicel} + N_{A,cellobiose} - 2.$$

The determined t-distribution value was compared to t-distributions corresponding to the varying degrees of freedom at different confidence levels. Relevant comparisons were made only with ratios for which the t-values were above the t-distributions at 95% confidence or better.

3.9. EmPAI analysis

EmPAI, which was shown to bear a roughly linear relationship to protein concentration, is defined as 10^{PAI} minus one, where PAI, the protein abundance index, is the ratio of the number of MS-observed peptides for a given protein over its theoretically observable peptides [138]. The unique peptide parent ions matched for a given protein were counted as its observed peptides. For theoretical peptides, a protein's *in silico* tryptic digest products (no missed cleavages, no PTMs) were generated within a mass window of 0 to 4000 Da using the MS-Digest tool at the ProteinProspector website (http://www1.ncifcrf.gov/ucsfhtml3.2/msdigest.htm). The relative hydrophobicities of the resulting peptide sequences were calculated using the Sequence Specific Retention Calculator available at http://hs2.proteome.ca/SSRCalc/SSRCalc.html [150]. Peptide retention times were predicted based on relative hydrophobicity and coefficients derived from our data set. These coefficients correspond to the slope and intercept of a plot of actual retention times against relative hydrophobicity values for a representative sample data set (Figure 11). Theoretical peptides were accepted within a retention time window of 12-68 min (the range of the regression line in Figure 11) and a mass window of 400-3500 Da (the same mass range used for SEQUEST searching). All emPAI values were normalized to that obtained for CipA, assuming that one CipA exists per cellulosome.

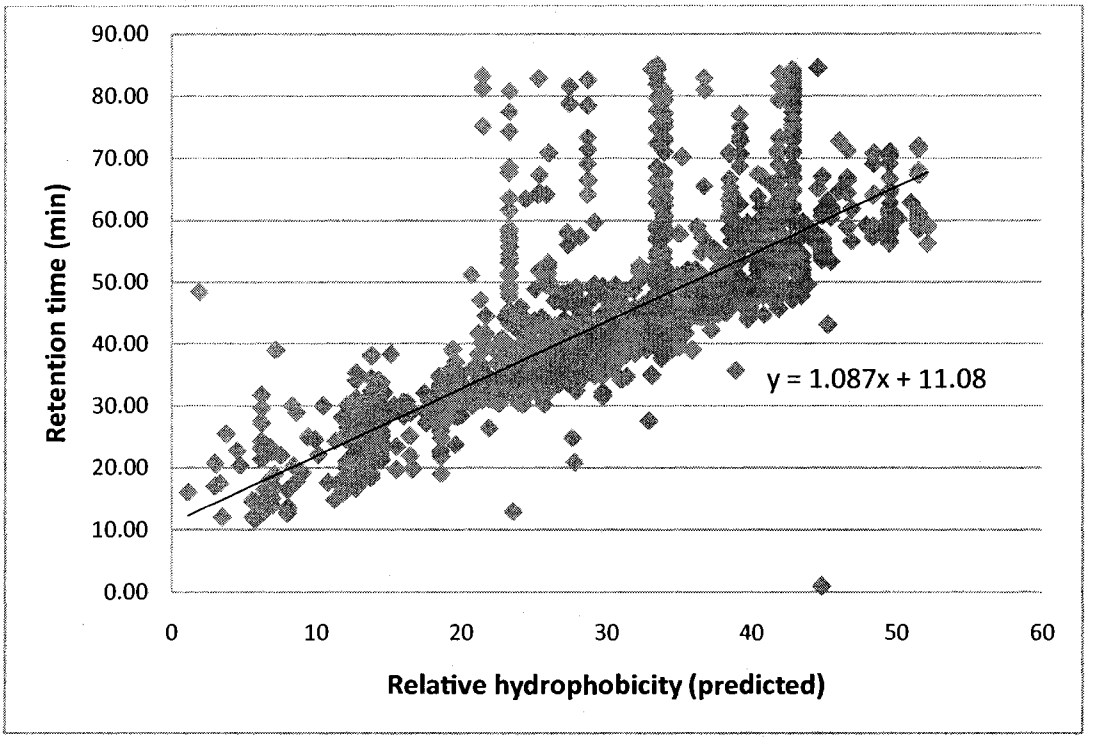

Figure 11. Determination of equation for predicting peptide retention times and determination of range for theoretically observable peptides used in emPAI analysis. Retention times for observed peptides are plotted against their calculated relative hydrophobicities. The slope of the linear regression line is used to calculate retention times for unobserved theoretical peptides.

The molar percentage of a docking subunit A per total docking subunits per CipA (thus not including anchor proteins) was calculated as

Eq. 5

$$\left(\frac{A}{CipA}\right)_{mol\%} = 100 \times \frac{\left(\frac{emPAI_A}{emPAI_{CipA}}\right)}{\Sigma\left(\frac{emPAI_{Doc1}}{emPAI_{CipA}}\right)}.$$

3.10. Enzymatic assays

Exoglucanase, endoglucanase and xylanase activities were tested for unmixed cellulosome preparations from cultures grown on Avicel, cellobiose, Avicel with XPM, or cellobiose with XPM. Activity against xylan and activity against carboxymethylcellulose (CMC) were determined by measuring the amount of reducing sugars released [151]. Bicinchoninic acid (BCA) was used as a detection reagent for Cu^{+1} which is formed upon reduction of Cu^{+2}. The complexing of two BCA molecules with one Cu^{+1} exhibits strong absorbance at 562 nm. The buffer solution for these assays was 53 mM succinate, pH 5.7, containing 2 mM $CaCl_2$. Substrate for xylanase activity was prepared by boiling a 0.5% (wt/vol) solution of xylan from birch wood (Sigma Aldrich) in water for 10 min, then centrifuging to remove insoluble xylan. CMC-4M (Megazyme) was dissolved in water to 0.5% (wt/vol). The total reaction volume was 80 µL: 40 µL of substrate (xylan or CMC-4M); 30 µL of 140 mM succinate, pH 5.7, buffer with 5mM $CaCl_2$; and 10 µL of enzyme dilution (Table 4). Enzyme preparations were diluted from 0.5 to 0.001. To determine concentrations of reducing sugars produced, standard curves were prepared using xylose and glucose in concentrations ranging from 0.005 to 1 mM. Both reactions were carried out in 0.5-mL microcentrifuge tubes, in a thermocycler block, for 15 min at 60°C. Ten µL of each reaction mixture were transferred to a clean

Table 4. Reaction mixtures for enzyme assays

	Xylanase	CMCase	PNPCase
Xylan (0.5 %, wt/vol) or xylose standard	40 μL		
CMC-4M (0.5 %, wt/vol) or glucose standard		40 μL	
PNPC (5 mM) or PNP standard			10 μL
140 mM succinate, pH 5.7; 5 mM $CaCl_2$	30 μL	30 μL	
100 mM succinate, pH 5.7			15 μL
Enzyme dilution (or water if standard or blank)	10 μL	10 μL	25 μL
Total volume	80 μL	80 μL	50 μL

tube, to which was added, on ice, BCA reagent (100 µL of a 1:1 mixture of a first solution containing 0.51 M Na_2CO_3, 0.29 M $NaHCO_3$, and 5 mM BCA, and a second solution containing 12 mM L-serine and 5 mM $CuSO_4$), followed by 90 µL of water. A 40-min incubation at 80°C followed to allow the colour to develop. Eighty µL of each reaction mixture was transferred to a 96-well plate, and absorbance was read at a wavelength of 562 nm. One unit (U) of xylanase or CMCase activity is defined as the amount of enzyme releasing 1 µmol of xylose or glucose equivalent from xylan or CMC-4M per min.

Activity against *p*-nitrophenyl-β-D-cellobioside (PNPC) was determined by measuring the release of *p*-nitrophenol (PNP), which itself exhibits strong absorbance at 410 nm at pH 10. The buffer solution for this assay was 30 mM succinate, pH 5.7. A 5 mM solution of PNPC in water was prepared as substrate. The total reaction volume was 50 µL: 10 µL of substrate; 15 µL of 100 mM succinate, pH 5.7, buffer; and 25 µL of enzyme dilution (Table 4). Enzyme preparations were diluted from 0.02 to 0.00125. A standard curve was prepared using PNP concentrations ranging from 0.0005 to 5 mM. Reactions were again carried out in 0.5-mL microcentrifuge tubes, in a thermocycler block, but for 60 min at 60°C. Fifty µL of a 1 M disodium carbonate solution was added to quench the reaction and raise the pH for colour development. Eighty µL of each reaction mixture were transferred to a 96-well plate, and absorbance was read at a wavelength of 410 nm. One unit of PNPCase activity is defined as the amount of enzyme releasing 1 µmol of PNP from PNPC per min.

Measurement of total protein for specific activity determination was done using bovine serum albumin as standard with the MicroBCA Protein Assay Kit (Pierce), which

functions on the same principle as the reducing sugar assay described above. Standard errors (SE) in the specific activity (s.a.) values reported were calculated from standard deviations from total protein assays performed in quadruplicate and enzyme assays performed in triplicate, as follows:

$$\textit{Eq. 6} \quad SE_{s.a.} = \sqrt{\left(s.a._1 \times \frac{SD_{total\ protein}}{total\ protein}\right)^2 + \left(s.a._2 \times \frac{SD_{total\ protein}}{total\ protein}\right)^2 + \left(s.a._3 \times \frac{SD_{total\ protein}}{total\ protein}\right)^2}$$

4. RESULTS

4.1. Fractionation of *C. thermocellum* protein

Total extracellular and cell-surface protein fractions (Figure 9, center) were obtained in order to assess their complexity and our ability and to isolate and detect cellulosomal protein. Protein fractions extracted from a cellobiose-grown *C. thermocellum* culture and separated by SDS-PAGE are shown in Figure 12. In the total extracellular protein fraction (Figure 12, lane A), the cellulosome scaffolding protein CipA appears above the 200 kDa mark; a smear between about 110 and 150 kDa likely corresponds to a glycosylated protein. When cellulose-binding protein was removed from the total extracellular protein fraction using the affinity digestion method, CipA disappeared although the smear remained (Figure 12, lane B). As expected, CipA was found to reside in the cellulose-binding fraction obtained via affinity digestion (Figure 12, lanes C and D).

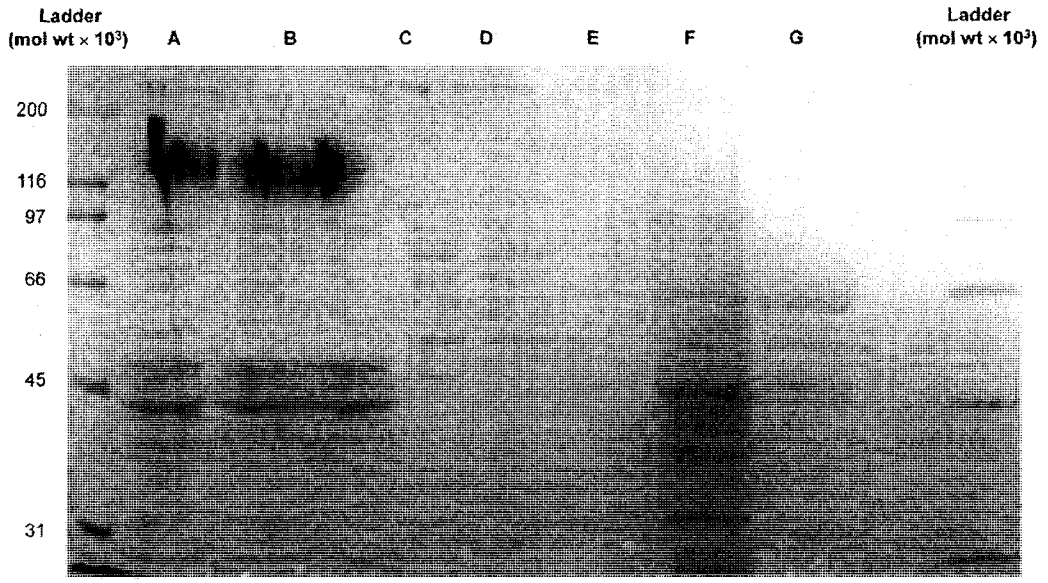

Figure 12. Extracellular protein fractions from *C. thermocellum* culture grown to late stationary phase on cellobiose (0.5%, wt/vol), separated by SDS-PAGE (6%), stained with Coomassie Blue. Lane A, total extracellular protein. Lane B, non cellulose binding protein. Lane C, cellulose binding (cellulosomal) protein fraction. Lane D, same as C with residual cellulose removed. Lane E, cell-surface protein released by treatment with lysozyme. Lane F, cell-surface protein released by treatment with urea (possible lysis). Lane G, cell-surface protein released by sonication. Mol wt markers shown at left and right.

The total extracellular protein fraction was digested with trypsin in-solution and then analyzed by one-dimensional nanoLC-ESI-MS. Only 5 of the proteins matched to the *C. thermocellum* database contained a predicted signal peptide cleavage site in their sequence such that they would be secreted into the supernatant (Table 5). For the analysis of cellulosomes, the fact that CipA was the only cellulosomal protein detected, and with such low percent amino acid coverage as compared to the other 4 proteins, suggests the importance of simplifying the protein fraction and/or adding a dimension of resolution, either prior to protein digestion (SDS-PAGE) or afterwards (strong cation exchange chromatography). Alternatively, it may be that the conditions used in the in-solution digest were not harsh enough to dissociate the cellulosome into its component proteins for the proteolysis step to be effective.

MS analysis of the gel bands containing the protein smear (Figure 12, lanes A and B) determined that it corresponds to a predicted 113-kDa protein (gi 125974833) with a possible (e = .006) SLH domain (pfam00395) and an immunoglobulin-like fold. The significance of the presence of this protein and of the other noncellulosomal proteins detected in the total extracellular protein fraction is addressed below in section 3.2.3.

Three methods were tested for releasing proteins from the cell surface. The mild sonication appears to be the most promising, as compared with treatment with lysozyme or urea. Judging from the number of proteins that appear on the gel (Figure 12, lane F), the urea treatment may have caused the cells to lyse.

Table 5. *C. thermocellum* proteins detected in the total extracellular protein fraction whose sequence also contained a signal peptide for secretion from the cell

GenInfo ID	(Putative) protein function	no. obs. un. pep.	P_{pro}	Score XC	Cov %AA	mol wt ($x10^3$)
125973535	Extracellular solute-binding protein, family 1	27	1.00E-30	246.3	57.52	50.0
125972914	Sugar ABC transporter	8	1.00E-30	60.3	25.55	33.9
125974833	Ig-related protein, SLH domain	52	1.78E-14	400.3	49.71	113.2
125975556	CipA scaffolding protein	4	8.04E-12	30.3	6.31	196.7
125973947	Short-chain dehydrogenase/reductase	2	1.80E-10	20.2	12.06	27.7

no. obs. un. pep, number of unique parent peptide ions matched (including different charge states, modifications)
P_{pro}, probability of finding a match as good or better than the observed match by chance
XC, protein cross correlation score calculated by SEQUEST
Cov %AA, percentage of amino acid coverage to the matched protein

4.2.1. Detection and relative abundance of cellulosomal proteins induced by Avicel or cellobiose

For investigation of substrate-induced changes to the cellulosomal subunit profile of *C. thermocellum*, cellulosome complexes were isolated from the supernatants of batch cultures grown to late stationary phase on either Avicel (the model substrate for crystalline cellulose) or cellobiose. Prior to cellulosome isolation, each culture was mixed with an equal volume of a ^{15}N-labelled Avicel-grown culture for quantitation at a later step. Purified cellulosomes were denatured and the components separated by SDS-PAGE. Proteins in the gel bands (Figure 13) were trypsin-digested and extracted for analysis.

In total, 41 cellulosomal proteins in the *C. thermocellum* database were detected between the two samples, 35 on Avicel (Table 6), 34 on cellobiose (Table 7), with 29 common to both samples. Thus, a similar number of subunits were detected in the two growth conditions. A total of 36 docking components were identified, including 16 subunits that have never been observed experimentally as components of the cellulosome. The specificity of the methodology is such that the matching of only two unique peptides to one protein out of the 3191 proteins in the *C. thermocellum* database resulted in a probability of at worst 10^{-5} that another protein could have been matched. The molecular weights of the proteins identified generally corresponded to the gel bands in which they were detected; deviations from this trend suggested possible proteolysis or glycosylation. The 17 new proteins identified in this study are indicated in Tables 6, 7 and 8 by shaded rows. The reference protein from Avicel-grown cells did not interfere with the identification of cellulosomal proteins from cellobiose-grown cells in the mixed sample as SEQUEST analysis could not identify ^{15}N-labelled peptides given the LC conditions

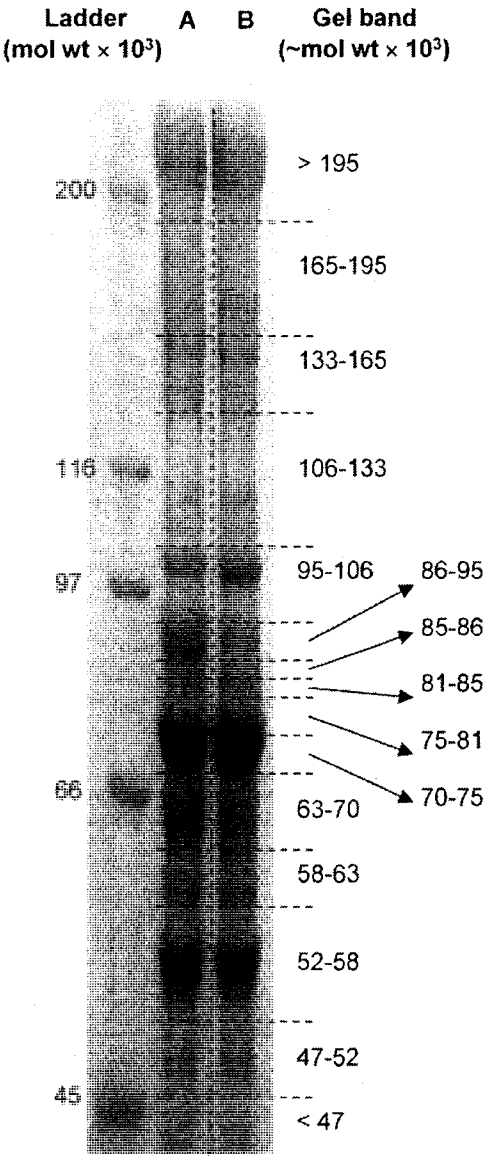

Figure 13. *C. thermocellum* cellulosomal protein separated by SDS-PAGE (6%), stained with Coomassie Blue. Lane A, 1:1 (vol/vol) mixture of unlabelled cellobiose-grown and ^{15}N-labelled Avicel-grown cellulosomes from late stationary phase, 170 μg total protein. Lane B, 1:1 (vol/vol) mixture of unlabelled Avicel-grown and ^{15}N-labelled Avicel-grown cellulosomes from late stationary phase, 170 μg total protein. Mol wt markers shown at left. At right, the approximate mol wt ranges for the division of the gel bands for trypsin digestion

Table 6. *C. thermocellum* Avicel-grown cellulosomal components identified by nanoLC-ESI-MS, ranked by emPAI. Proteins shaded in grey are those that have never been observed experimentally in purified cellulosomes previous to this study

GenInfo Identifier	Protein Name	(Putative) function or fold(s)	No. obs Un. Pep	emPAI	emPAI /CipA	DocI/CipA (mol%)	P_{pro}	Score XC	Cov %AA	mol wt ($x10^3$)	mol wt ($x10^3$) gel band	Ref.
125975556	CipA	Scaffolding protein	42	5.92	1.00		2.2E-12	378	34	196.7	> 195	[68]
125972933	CelK	Exoglucanase (GH9)	39	4.12	0.70	11.0	2.0E-12	350	35	100.6	95-106	[98]
125974579	CelS	Exoglucanase (GH48)	29	3.56	0.60	9.4	6.9E-10	240	32	83.5	75-81	[68]
125973097	CelR	Endoglucanase (GH9)	28	3.19	0.54	8.5	9.3E-11	250	31	82.1	81-86	[91]
125975557	OlpB	Cell-surface anchoring protein	27	2.75	0.47		3.4E-13	210	26	248.0	165-195	[152]
125973339	--	GH5	15	2.59	0.44	6.9	1.3E-09	140	25	63.0	59-63	
125972791	CelA	Endoglucanase (GH8)	14	2.46	0.41	6.4	1.1E-08	120	22	52.6	52-59	[153]
125973315	CelE	Endoglucanase (GH5), CE2	24	2.24	0.38	6.0	9.1E-13	200	32	90.2	86-95	[154]
125973142	CelJ	Endoglucanase (GH9), GH44	42	2.16	0.37	5.8	2.4E-13	390	24	178.0	165-195	[153]
125974342	XynC	Xylanase (GH10)	18	1.57	0.26	4.1	2.0E-10	160	24	69.5	81-95	[153]
125974464	XynZ	Xylanase (GH10), CE1	18	1.57	0.26	4.1	2.0E-09	150	20	92.2	63-70	[91]
125972934	CbhA	Exoglucanase (GH9)	30	1.51	0.26	4.1	3.2E-11	280	22	137.0	133-165	[155]
125975294	CelT	Endoglucanase (GH9)	14	1.29	0.22	3.4	6.0E-11	120	21	68.5	59-70	[156]
125975353	CelG	Endoglucanase (GH5)	9	1.15	0.20	3.1	1.1E-07	90	16	63.2	81-85	[99]
125975452	XynA	Xylanase (GH11), CE4	12	1.15	0.20	3.1	1.6E-08	100	14	74.4	47-52	[157]
125973912	XghA	Xyloglucanase (GH74)	21	1.13	0.19	3.0	3.4E-10	180	22	92.3	86-95	[91]
125973263	--	GH9	15	0.94	0.16	2.5	7.4E-08	140	17	82.1	85-95	
125973062	CelF	Endoglucanase (GH9)	12	0.90	0.15	2.4	3.3E-08	120	19	82.0	81-85	[115]
125973254	--*	Cellulosome anchoring protein	21	0.82	0.14		5.6E-09	178	19	140.5	133-165	
125974678	--	GH5	11	0.71	0.12	1.9	4.0E-10	100	9.3	103.1	106-133	
125972735	LicB*	Lichenase (GH16)	4	0.62	0.11	1.7	4.1E-07	28.2	8.4	37.9	< 47	[158]
125975293	ManA*	Mannanase (GH26)	7	0.61	0.10	1.6	7.5E-10	70.2	19	67.0	63-70	[159]
125973143	CelQ*	Endoglucanase (GH9)	9	0.58	0.10	1.6	1.0E-13	70.2	14	79.8	70-81	[108]
125972556	--	GH26	5	0.53	0.09	1.4	1.2E-07	40.2	5.8	67.3	63-70	
125973055	CelB	Endoglucanase (GH5)	5	0.53	0.09	1.4	6.1E-08	50.2	7.6	63.9	63-70	[1]
125975558	Orf2p	Cell-surface anchoring protein	7	0.50	0.08		5.9E-08	70.2	15	74.9	86-95	[121]
125972796	--*	GH9	5	0.49	0.08	1.3	3.7E-06	50.2	11	62.6	59-63	
125975243	--	GH9	8	0.44	0.07	1.1	6.7E-06	80.2	9.3	80.2	85-86	
125972926	--	GH5	3	0.33	0.06	0.9	1.7E-06	30.1	6.3	59.9	59-63	
125972567	CelN*	Endoglucanase (GH9)	4	0.27	0.05	0.8	3.1E-06	40.2	5.7	82.1	85-86	[91]
125973343	CelD	Endoglucanase (GH9)	4	0.23	0.04	0.6	6.2E-06	30.2	5.2	72.4	59-70	[160]
125972954	--	GH9	4	0.21	0.04	0.6	7.3E-06	40.2	2.9	89.4	95-106	
125972792	ChiA	Chitinase (GH18)	2	0.20	0.03	0.5	8.9E-08	20.2	4.8	55.4	47-52	[161]
125973914	--	GH53	2	0.17	0.03	0.5	5.5E-07	20.1	6.0	47.0	< 47	
125973158	--*	Endopygalactorunase	2	0.16	0.03	0.5	7.0E-07	20.2	5.2	64.5	59-63	

GH, glycoside hydrolase family; CE, carbohydrate esterase family; *, found only in Avicel sample; no. obs. un. pep, number of unique parent peptide ions matched (including different charge states, modifications); emPAI, exponentially modified protein abundance index; /CipA, normalized to the value obtained for CipA; DocI/CipA (mol%), molar percentage per CipA for dockerin I-containing subunits, calculated as $100 \times (\text{emPAI/CipA})/[\Sigma(\text{emPAI/CipA})_{\text{docking subunits}}]$; P_{pro}, probability of finding a match as good or better than the observed match by chance; XC, protein cross correlation score calculated by SEQUEST; Cov %AA, percentage of amino acid coverage to the matched protein

Table 7. *C. thermocellum* cellobiose-grown cellulosomal components identified by nanoLC-ESI-MS, ranked by emPAI. Proteins shaded in grey are those that have never been observed experimentally in purified cellulosomes previous to this study

GenInfo Identifier	Protein Name	(Putative) function or fold(s)	no. obs un. pep	emPAI	emPAI /CipA	Doc1/CipA (mol%)	P_{pro}	Score XC	Cov %AA	mol wt ($x10^3$)	mol wt ($x10^3$) gel band	Ref.
125974464	XynZ	Xylanase (GH10), CE1	25	2.70	1.25	14.0	4.0E-13	200	31	92.2	86-95	[91]
125975556	CipA	Scaffolding protein	25	2.16	1.00		1.5E-10	230	30	196.7	> 195	[68]
125975452	XynA	Xylanase (GH11), CE4	17	1.97	0.91	10.2	1.9E-11	170	31	74.4	59-85	[157]
125974342	XynC	Xylanase (GH10)	16	1.31	0.61	6.8	8.1E-11	160	20	69.5	63-70	[153]
125972791	CelA	Endoglucanase (GH8)	9	1.22	0.56	6.3	7.2E-10	90.2	22	52.6	52-59	[153]
125973912	XghA	Xyloglucanase (GH74)	22	1.21	0.56	6.3	1.6E-10	210	27	92.3	86-95	[91]
125973339	--	GH5	9	1.15	0.53	5.9	3.2E-05	80.2	13	63.0	59-63	
125972933	CelK	Exoglucanase (GH9)	18	1.12	0.52	5.8	6.7E-09	180	21	100.6	95-106	[98]
125973315	CelE	Endoglucanase (GH5), CE2	14	0.99	0.46	5.1	1.1E-12	120	20	90.2	86-95	[154]
125972556	--	GH26	8	0.98	0.45	5.0	1.4E-08	60.2	9.5	67.3	63-70	
125973055	CelB	Endoglucanase (GH5)	7	0.82	0.38	4.2	6.4E-07	70.2	15	63.9	63-70	[1]
125975557	OlpB	Cell-surface anchoring protein	11	0.71	0.33		1.2E-10	90.2	14	248.0	165-195	[152]
125974678	--	GH5	10	0.63	0.29	3.2	7.6E-07	88.2	11	103.1	106-133	
125975073	XynD*	Xylanase (GH10)	7	0.58	0.27	3.0	4.7E-07	70.2	15	71.6	70-75	[91]
125975353	CelG	Endoglucanase (GH5)	5	0.53	0.25	2.8	3.9E-08	50.1	9.2	63.2	63-70	[99]
125973097	CelR	Endoglucanase (GH9)	8	0.51	0.23	2.6	9.6E-08	90.2	8.6	82.1	81-85	[91]
125973062	CelF	Endoglucanase (GH9)	7	0.45	0.21	2.3	1.3E-05	70.2	6.4	82.0	81-85	[115]
125972934	CbhA	Exoglucanase (GH9)	11	0.40	0.19	2.1	5.5E-07	108	8.6	137.0	133-165	[155]
125972926	--	GH5	3	0.33	0.15	1.7	1.6E-07	30.1	6.5	59.9	59-63	
125973786	--*	GH43	5	0.32	0.15	1.7	2.9E-06	50.1	7.8	74.5	63-75	
125975243	--	GH9	6	0.31	0.14	1.6	1.7E-06	60.2	6.5	80.2	85-86	
125975294	CelT	Endoglucanase (GH9)	4	0.27	0.12	1.3	3.6E-05	40.2	6.5	68.5	59-63	[156]
125973263	--	GH9	5	0.25	0.11	1.2	8.9E-08	50.2	6.6	82.1	85-86	
125974579	CelS	Exoglucanase (GH48)	4	0.23	0.11	1.2	9.8E-08	40.1	5.4	83.5	75-81	[68]
125972792	ChiA	Chitinase (GH18)	2	0.20	0.09	1.0	7.3E-08	20.1	2.5	55.4	47-52	[161]
125973142	CelJ	Endoglucanase (GH9), GH44	6	0.18	0.08	0.9	1.3E-07	54.2	3.4	178.0	165-195	[153]
125973914	--	GH53	2	0.17	0.08	0.9	5.0E-05	20.1	6.0	47.0	< 47	
125972954	--	GH9	3	0.15	0.07	0.8	6.3E-05	28.2	3.3	89.4	95-106	
125972540	--*	GH43, α-L-arabinofuranosidase B	3	0.15	0.07	0.8	1.0E-05	30.2	4.1	79.0	63-75	
125975558	Orf2p	Cell-surface anchoring protein	2	0.12	0.06		5.5E-05	20.2	1.9	74.9	85-95	[121]
125973343	CelD	Endoglucanase (GH9)	2	0.11	0.05	0.6	2.8E-06	20.2	1.8	72.4	59-63	[160]
125973822	SdbA*	Cell-surface anchoring protein	2	0.10	0.05		3.9E-05	20.2	4.6	68.6	63-70	[162]
125974626	--*	GH30, α-L-arabinofuranosidase B	2	0.09	0.04	0.4	1.4E-06	20.2	2.7	110.6	106-133	
125973429	XynY*	Xylanase (GH10), CE1	2	0.07	0.03	0.3	9.5E-06	20.2	1.2	119.6	106-133	[36]

GH, glycoside hydrolase family; CE, carbohydrate esterase family; *, found only in cellobiose sample; no. obs. un. pep, number of unique parent peptide ions matched (including different charge states, modifications); emPAI, exponentially modified protein abundance index; /CipA, normalized to the value obtained for CipA; Doc1/CipA (mol%), molar percentage per CipA for dockerin I-containing subunits, calculated as $100 \times (\text{emPAI/CipA})/[\Sigma(\text{emPAI/CipA})_{\text{docking subunits}}]$; P (pro), probability of finding a match as good or better than the observed match by chance; XC, protein cross correlation score calculated by SEQUEST; Cov %AA, percentage of amino acid coverage to the matched protein

and MS parameters applied. This was tested in an earlier experiment (data not shown), where ^{15}N-labelled cellulosomes were isolated independently and analysed by nanoLC-ESI-MS. No proteins were identified using SEQUEST and the same criteria as described above.

The emPAI method [138] was used to relate the number of unique peptides matched to a protein to the relative abundance of that protein in each sample. While attempts to standardize the emPAI method on our system revealed a divergence from linearity at higher concentrations such that higher abundance proteins would be underestimated, it nevertheless supplies a basis for informed analysis as to the abundance of particular proteins per cellulosome preparation. Since the affinity digestion method used to isolate cellulosomes pulls the complex down 'by the CipA', all relative abundance values (emPAI and RelEx below) were normalized to that obtained for CipA. This provided a protein-per-CipA basis for comparison between samples.

There are significant differences in the relative abundances of docking subunits per CipA between the two data sets as per molar percentage calculated from emPAI values. Exoglucanases accounted for a total molar percentage of 24.4% of the total moles per CipA of all docking subunits detected when cells were grown on Avicel, but only 9.2% when cells were grown on cellobiose. The molar percentage of CelS dropped from 9.4% on Avicel to 1.2% on cellobiose, while GH9 exoglucanases CelK and CbhA changed from 11.0 to 5.8% and 4.1 to 2.1%, respectively. Components with known endoglucanase activity accounted for a total molar percentage of 40.0% when cells were grown on Avicel, but this decreased to 26.1% on cellobiose. In total, GH9 cellulases decreased from 43.6% on Avicel to 19.2% on cellobiose; whereas enzymes containing a

GH5 domain increased slightly from 20.2% on Avicel to 23.0% on cellobiose. The GH5 fold is predominantly associated with cellulases, but it has also been linked to hemicellulolytic activity [21]. A new GH5 enzyme (gi 125973339) was detected among the most abundant catalytic subunits in both samples (6.9% on Avicel, 5.9% on cellobiose). It has a predicted mass of 63.0 kDa and migrated similarly as known proteins CelB and CelT by SDS-PAGE when isolated from cells grown both conditions (Tables 6 and 7); the overlap with these proteins might explain why it was not identified previously. Overall, the molar percentage of hemicellulases increased from 19.9% on Avicel to 50.3% on cellobiose. Docking subunits with xylanase activity accounted for a total of 11.3% of all docking subunits detected when cells were grown on Avicel, but their contribution increased to 34.3% when cells were grown on cellobiose. Other hemicellulases accounted for a total molar percentage of 8.6% on Avicel and 15.1% on cellobiose. GH9 cellulases were the most abundant group of enzymes per CipA when cells were grown on Avicel, while hemicellulases were the most abundant group on cellobiose.

Other notable differences between the two samples concern the 13 components detected exclusively in one sample but not the other. Detected only in Avicel-grown cellulosomes were GH9 endoglucanases CelN and CelQ; the GH16 lichenase LicB; the GH26 mannanase ManA; a new GH9 cellulase; a new subunit with putative endopygalactorunase activity; and a new cell-surface anchor protein predicted to have the same number of type II cohesin domains as OlpB but no SLH domain. XynD and XynY, both with GH10 xylanase activity, were detected exclusively in cellobiose-grown cellulosomes, along with cell-surface anchoring protein SdbA, a new bifunctional

GH30/α-L-arabinofuranosidase B hemicellulase, a new GH43 glycosidase, and a new bifunctional GH43/α-L-arabinofuranosidase B glycosidase.

4.2.2. Relative differences in abundance of cellulosomal components induced by Avicel or cellobiose

Simultaneous quantitative differences in the expression of all but four cellulosomal components common to both Avicel and cellobiose were measured by means of metabolically ^{15}N-labelled peptides as internal standards. While emPAI supplied a means of determining the relative abundance of proteins in a given sample, RelEx provides a highly reliable way to compare the amount of a particular protein present in two samples. Sample-to-reference ratios were determined separately for Avicel- and cellobiose-grown cellulosomes, and the ratio of ratios represented the fractional difference between proteins grown on either substrate. Normalization of ratio values to that obtained for the scaffoldin protein CipA allowed for comparison of changes in protein expression per cellulosome complex. That the average ratio of unlabelled Avicel-grown protein to ^{15}N-labelled protein was 1.23 with a standard deviation of 0.29 (Table 8) suggests that our methodology was accurate (and precise) at determining ratios between cellulosomal proteins from two separate samples.

From the total of 29 proteins found in both samples, RelEx was able to determine a ratio of sample-to-reference for 25 protein pairs, given the S/N and correlation filters adopted (Table 8). The null hypothesis was rejected for all but four of these, for which it was determined that $p \geq .05$. There was no significant change in expression for these four proteins: two new GH9 cellulases and two hemicellulases, ChiA and a new GH53

Table 8. Fractional differences in expression of *C. thermocellum* Avicel-grown cellulosomal components relative to cellobiose-grown components by RelEx, ranked by *p*-value, normalized to CipA. Proteins shaded in grey are those that have never been observed experimentally previous to this study

GenInfo Identifier	Protein name or type	^{14}N Avicel / ^{15}N Avicel ratio1	SD	no. pep	^{14}N cellobiose / ^{15}N Avicel ratio2	SD	no. pep	^{14}N Avicel / ^{14}N cellobiose ratio1/ratio2	*p*-value	/CipA	SE	Overall change on Avicel
125975557	OlpB	1.404	0.299	80	0.134	0.016	7	10.453	< 0.0001	2.063	0.672	Increase
125975556	CipA	1.574	0.244	179	0.311	0.047	108	5.067	< 0.0001	1.000	0.305	Increase
125973339	GH5	1.358	0.138	14	0.314	0.031	10	4.320	< 0.0001	0.853	0.220	None
125974678	GH5	1.171	0.131	7	0.751	0.074	6	1.559	< 0.0001	0.308	0.081	Decrease
125972791	CelA	0.939	0.112	21	0.659	0.067	9	1.426	< 0.0001	0.281	0.075	Decrease
125973142	CelJ	1.959	0.220	48	0.121	0.019	3	16.191	< 0.0001	3.195	0.926	Increase
125972933	CelK	1.776	0.353	76	0.158	0.015	14	11.214	< 0.0001	2.213	0.680	Increase
125975294	CelT	1.338	0.244	14	0.197	0.011	3	6.791	< 0.0001	1.340	0.386	None
125973062	CelF	1.041	0.250	18	0.161	0.007	5	6.452	< 0.0001	1.273	0.415	None
125972934	CbhA	1.030	0.180	39	0.166	0.019	6	6.211	< 0.0001	1.226	0.369	None
125973097	CelR	1.441	0.325	27	0.252	0.032	7	5.719	< 0.0001	1.129	0.380	None
125974342	XynC	1.105	0.102	21	0.528	0.030	13	2.094	< 0.0001	0.413	0.100	Decrease
125974464	XynZ	1.043	0.264	32	4.323	1.244	48	0.241	< 0.0001	0.048	0.021	Decrease
125975452	XynA	1.095	0.25	27	2.144	0.684	32	0.511	< 0.0001	0.101	0.045	Decrease
125974579	CelS	0.932	0.138	81	0.028	0.006	4	33.274	< 0.0001	6.567	2.171	Increase
125973912	XghA	1.035	0.151	24	1.662	0.332	33	0.623	< 0.0001	0.123	0.040	Decrease
125975353	CelG	0.947	0.245	10	0.333	0.035	4	2.842	0.0004	0.561	0.198	Decrease
125972556	GH26	0.871	0.090	3	1.748	0.197	5	0.499	0.0004	0.098	0.026	Decrease
125973315	CelE	1.032	0.187	23	0.736	0.205	9	1.401	0.0005	0.277	0.110	Decrease
125973263	GH9	1.438	0.408	12	0.546	0.061	4	2.633	0.0008	0.520	0.194	Decrease
125973055	CelB	0.996	0.080	4	1.320	0.231	6	0.754	0.0291	0.149	0.043	Decrease
125972954	GH9	1.565	0.245	2	0.938	0.073	2	1.669	0.0742	0.329	0.091	None
125975243	GH9	0.978	0.185	8	1.110	0.102	6	0.881	0.1426	0.174	0.052	None
125972792	ChiA	1.199	0.054	2	0.774	0.253	2	1.548	0.1463	0.306	0.121	None
125973914	GH53	1.372	0.257	2	1.007	0.198	2	1.362	0.2528	0.269	0.093	None

GH, glycoside hydrolase family; *p*-value, probability that the null hypothesis is true, based on 2-tailed Student t-test of ratio1 versus ratio2; /CipA, normalized to the value obtained for CipA; SE, standard error calculated using the simple quotient rule of error propagation, where a protein's ratio on Avicel, its ratio on cellobiose, CipA's ratio on Avicel, and CipA's ratio on cellobiose were considered random and independent

subunit, whether obtained from Avicel- or cellobiose-grown cells. Proteins for which significant differences were observed are represented visually over a logarithmic scale in Figure 14.

The grouping of proteins by structural function or enzymatic activity revealed several trends. Cell-surface anchoring protein OlpB demonstrated higher expression during growth on Avicel than on cellobiose (Table 8), suggesting an increased anchoring requirement for a greater number of cellulosomes. Expression of exoglucanases was either higher in Avicel-grown cellulosomes or showed no change as compared to growth on cellobiose. As expected, based on the results of a previous study, cellobiohydrolase CelS showed the greatest difference in favour of growth on Avicel of any docking enzyme. GH9 endoglucanases either demonstrated higher expression on Avicel (CelJ) than on cellobiose, or exhibited no significant change between the two substrates (CelT, CelF, CelR). On the other hand, GH8 endoglucanase CelA and GH5 endoglucanases (CelB, CelE, CelG) showed lower expression on Avicel than on cellobiose. One new enzyme from each of GH9 and GH5 demonstrated higher expression in cells grown on cellobiose. All hemicellulases compared displayed higher expression per cellulosome when cells were grown on cellobiose.

4.2.3. Non-cellulosomal proteins detected in Avicel- or cellobiose-grown cells

Four non-cellulosomal proteins with signal peptides for secretion were detected (not shown in Tables 6 or 7). The GH9 endoglucanase CelI (gi 125972564) was detected in the cellobiose cellulosome sample [110]. It was identified by two unique peptides. From the Avicel-grown sample only, three unique peptides were matched to a predicted

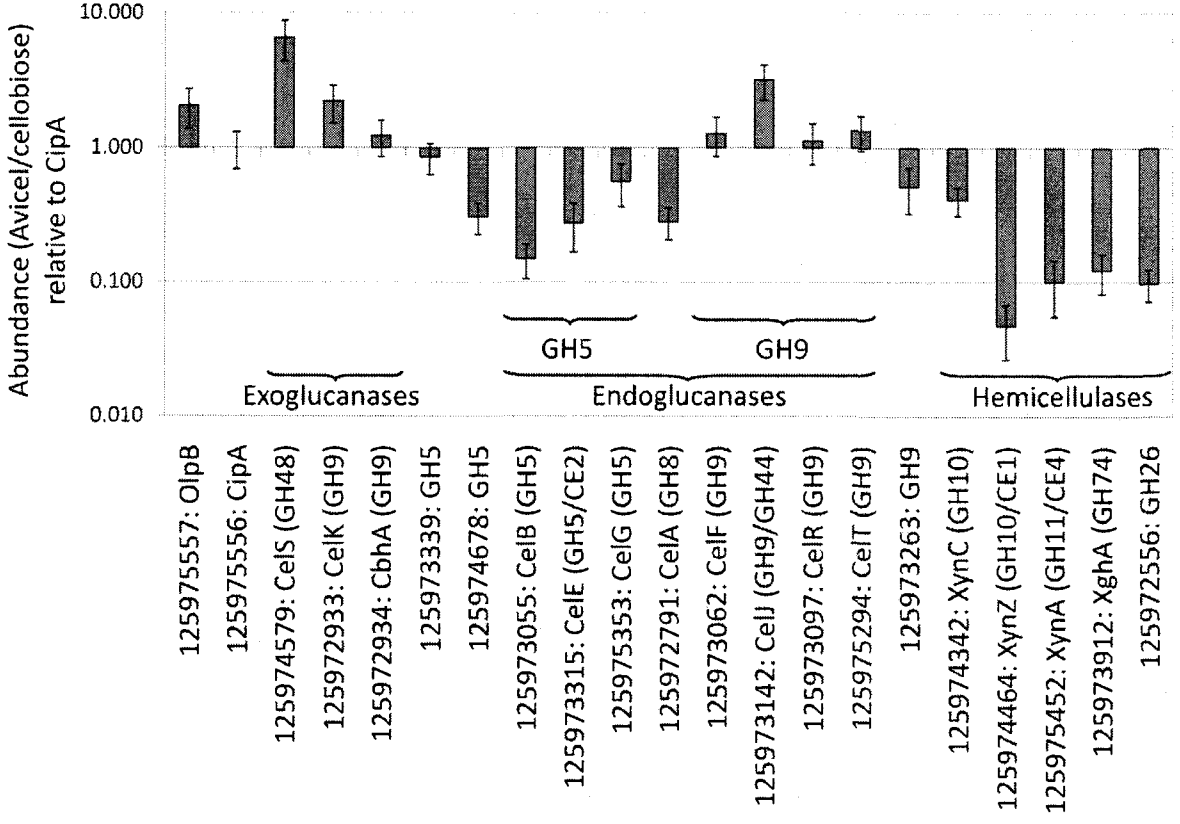

Figure 14. Fractional differences in expression of *C. thermocellum* Avicel-grown cellulosomal components relative to cellobiose-grown components by RelEx, normalized to CipA, over logarithmic scale. Docking components grouped by function and activity. GH, glycoside hydrolase family. CE, carbohydrate esterase family. Only proteins passing null hypothesis with $p < 0.05$ shown. Columns rising above 1 represent proteins determined to have greater expression in Avicel-grown sample. Columns falling below 1 represent proteins with higher expression in cellobiose-grown sample. Error bars traversing 1 signify no change in expression between the two samples.

34-kDa protein (gi 125972914) with similarity (e = 3E-32) to RbsB (COG1879), a ribose-binding protein in *Escherichia coli*. This protein also has a lipid attachment site to anchor it to the membrane. In both Avicel- and cellobiose-grown cellulosome preparations, 17 and 10 unique peptides, respectively, matched to a predicted 50-kDa protein (gi 125973535) with similarity (e = 1E-42) to UgpB (COG1653), a periplasmic glycerol-3-phosphate-binding protein in *E. coli*. Finally, seven unique peptides from both samples were matched to a predicted 113-kDa protein (gi 125974833) with a possible SLH domain for anchoring it to the cell wall, and also an immunoglobulin-like fold, which may behave like a carbohydrate binding domain. This protein had been recently observed in the cell membrane fraction [36], and its migration pattern by SDS-PAGE suggests it may be glycosylated. All three of the latter proteins were observed in considerable abundance (at least 25% amino acid coverage) in the total extracellular protein fraction from cells grown on cellobiose (Table 5, section 4.1). Their high abundance and, more particularly, the presence in each of them of a possible carbohydrate binding domain point to the possibility that these proteins are contaminants of the cellulosome preparations, consistently co-purifying with cellulosome-cellulose complexes. This possibility does not, however, preclude the alternative: that they may in fact be specifically associated with these complexes and play roles in secondary cellulosomal product-related function, perhaps in the uptake of cellodextrins in the manner of RbsB from *Bacillus subtilis* [163] or MalX from *Streptococcus pneumoniae* [164], both lipoproteins involved in sugar transport in Gram-positive bacteria.

4.3.1. Comparison of cellulosomes from cells grown in medium containing xylan, pectin and locust bean gum

A subsequent experiment compared *C. thermocellum* cellulosomes grown on four sets of substrates: cellobiose; cellobiose with xylan (X), pectin (P), and locust bean gum (M); Avicel; and Avicel with XPM. X, P and M were added all together with the expectation that xylanase, pectinase and mannanase expression would increase, such that yet more novel cellulosomal components could be detected. The Avicel and cellobiose samples were included as controls based on our previous findings. Each of the four cultures was grown to stationary phase and then mixed in equal volume with a ^{15}N-labelled reference culture, this time grown on Avicel with XPM. As before, cellulosomes were isolated from each mixture, separated by SDS-PAGE (Figure 15), digested with trypsin, and then separated and detected by nanoLC-ESI-MS for subsequent emPAI and RelEx analysis.

Fourteen docking proteins not observed in the previous experiment (described in section 3.2) were detected between the four samples, but in low abundance as per emPAI per CipA (Table 9). Among these was the GH5 exoglucanase CelO, which was detected only on Avicel with XPM. CseP was detected in both the Avicel and cellobiose samples, whereas PinA was detected in the latter only. A new GH30 docking component was detected in all but the Avicel sample, and in quite high abundance on cellobiose with XPM. A new pectate lyase (gi 125975431) was detected in all 4 samples. One new lipase (gi 125973316) was detected in all but the Avicel sample, while another lipase (gi 125975619) was detected in only the cellobiose sample containing XPM.

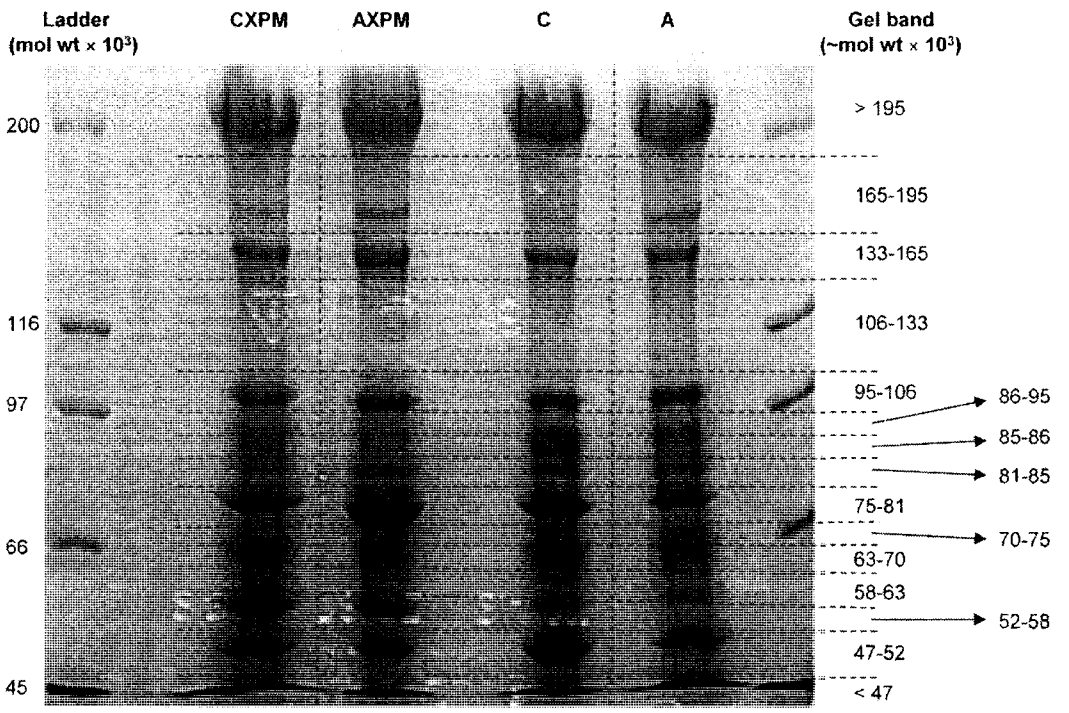

Figure 15. *C. thermocellum* cellulosomal protein from sample-reference culture mixtures with cells grown on Avicel or cellobiose, with and without XPM, separated by SDS-PAGE (6%), stained with Coomassie Blue. Lane CXPM, 1:1 (vol/vol) mixture of reference cellulosomes with cellulosomes grown on cellobiose supplemented with xylan, pectin and locust bean gum. Lane AXPM, 1:1 (vol/vol) mixture of reference cellulosomes with cellulosomes grown on Avicel supplemented with xylan, pectin and locust bean gum. Lane C, 1:1 (vol/vol) mixture of reference cellulosomes with cellulosomes grown on cellobiose. Lane A, 1:1 (vol/vol) mixture of reference cellulosomes with cellulosomes grown on Avicel. Approximately 150 μg total protein per sample lane. Mol wt markers shown at left. At right, the approximate mol wt ranges for the division of the gel bands for trypsin digestion.

Table 9. Comparison of relative cellulosome component abundances per CipA per sample, as determined by emPAI, for cellulosomes grown on Avicel (A) and cellobiose (C) with or without xylan (X), pectin (P), and locust bean gum (M), organized by protein function or fold. Shaded rows indicate proteins not detected in the previous experiment described in section 4.2

	GenInfo ID	Protein name and (putative) function or fold(s)	mol wt ($\times 10^3$)	C	CXPM	A	AXPM
Structural proteins	125973254	cell-surface anchor	140.5	0.17	0.08	0.15	0.42
	125973822	SdbA cell-surface anchor	68.6	0.23	0.21	0.03	0.05
	125975556	CipA scaffoldin	196.7	1.00	1.00	1.00	1.00
	125975557	OlpB cell-surface anchor	248.0	0.82	0.51	0.90	0.89
	125975558	Orf2p cell-surface anchor	74.9	0.28	0.23	0.17	0.17
Exoglucanases	125974633	CelO exoglucanase (GH5)	75.3				0.07
	125972933	CelK cellobiohydrolase (GH9)	100.6	0.70	0.90	0.43	1.96
	125972934	CbhA cellobiohydrolase (GH9)	137.0	0.46	0.44	0.35	0.56
	125974579	CelS cellobiohydrolase (GH48)	83.5	1.27	1.47	0.60	3.38
			Total	**2.43**	**2.81**	**1.37**	**5.97**
		Total docking %	**per CipA**	**12.14**	**9.59**	**15.39**	**43.66**
Endoglucanases	125973055	CelB endoglucanase (GH5)	63.9	1.38	1.45	0.22	0.38
	125973315	CelE endoglucanase (GH5), CE2	90.2	0.26	0.16	0.24	0.26
	125975353	CelG endoglucanase (GH5)	63.2	0.82	0.63	0.40	0.47
	125972791	CelA endoglucanase (GH8)	52.6	3.41	5.62	2.00	0.43
	125972567	CelN endoglucanase (GH9)	82.1	0.19	1.22	0.04	0.07
	125973062	CelF endoglucanase (GH9)	82.0	0.25	0.79	0.11	0.13
	125973097	CelR endoglucanase (GH9)	82.1	0.77	1.94	0.82	0.69
	125973142	CelJ endoglucanase (GH9), GH44	178.0	0.59	0.26	0.42	0.44
	125973143	CelQ endoglucanase (GH9)	79.8	0.10	0.07	0.02	0.36
	125973343	CelD endoglucanase (GH9)	72.4	0.14	0.08		0.07
	125975294	CelT endoglucanase (GH9)	68.5	0.35	0.34	0.08	0.35
			Total	**8.27**	**12.57**	**4.36**	**3.65**
		Total docking %	**per CipA**	**41.28**	**42.92**	**48.86**	**26.70**
Uncharacterized GH5 and GH9	125972926	GH5	59.9	0.31	0.08	0.07	
	125973339	GH5	63.0	0.46	5.23	0.17	1.59
	125974678	GH5	103.1	0.24	0.06		0.04
	125972796	GH9	62.6	0.42	0.15	0.30	0.17
	125972954	GH9	89.4		0.08		0.09
	125973263	GH9	82.1	0.10	0.23	0.18	0.14
	125975242	GH9	109.0	0.02			
	125975243	GH9	80.2	0.16	0.11	0.10	
Xylanases	125973429	XynY xylanase (GH10)	119.6	0.03			
	125974342	XynC xylanase (GH10)	69.5	1.92	3.16	0.44	0.46
	125974464	XynZ xylanase (GH10), CE1	92.2	1.27	0.63	0.79	0.22
	125975073	XynD xylanase (GH10)	71.6	0.75	0.44	0.19	0.10
	125975452	XynA xylanase (GH11), CE4	74.4	0.67	1.38	0.45	0.23
			Total	**4.64**	**5.61**	**1.87**	**1.01**
		Total docking %	**per CipA**	**23.16**	**19.16**	**20.98**	**7.39**
Other hemicellulases	125972735	LicB lichenase (GH16)	37.9		0.30		
	125975291	GH16	147.8	0.05			
	125972792	ChiA chitinase (GH18)	55.4	0.06			
	125972556	GH26	67.3	0.53	0.10		
	125975293	ManA mannanase (GH26)	67.0	0.25	0.27	0.06	0.25
	125975491	GH30	70.4	0.13	0.69		0.22
	125974626	GH30, α-L-arabinofuranosidase B	110.6			0.03	
	125972540	GH43, α-L-arabinofuranosidase B	79.0	0.05			
	125973179	Galactan 1,3-β-galactosidase (GH43)	63.9	0.04		0.05	0.04
	125973786	GH43	74.5	0.22	0.20	0.05	0.14
	125973914	Arabinogalactan endo-1,4-galactosidase (GH53)	47.0	0.08			0.05
	125973912	XghA xyloglucanase (GH74)	92.3	0.90	0.22	0.26	0.15
	125973816	Lipase	58.1	0.23	0.11		0.07
	125975431	Pectate lyase	59.4	0.05	0.15	0.05	0.05
	125975819	Lipolytic enzyme	91.2		0.09		
Other docking components	125972565	CttP sporangesome assembly	61.3	0.08	0.09		
	125972714	PntA serine protease inhibitor	67.8	0.10			0.04
	125972761	Unknown cellulosome enzyme	117.5	0.02			
	125972768	Integrin alpha chain	89.3		0.06		
	125973158	Endopygalactorunase	64.5	0.11			
	125973247	Unknown cellulosome enzyme	56.7	0.08			
	125975610	Unknown cellulosome enzyme	46.8		0.07		

The different chromatographic conditions used – longer run time, shallower gradient, and larger-bore column (the use of a 180-μm instead of a 75-μm internal diameter column corresponds to about a 6-fold drop in mass sensitivity) – resulted in a lower number of theoretically observable peptides for this experiment as compared to the previous one (Table 11). In spite of this, emPAI, which relies on the number of unique peptides observed, can still be used to glean semi-quantitative information, although the emPAI results should be considered with caution. Protein bands between 165-195 kDa and most pronounced in each of the Avicel and Avicel with XPM sample lanes of the SDS-PAGE gel (Figure 15) appeared to correspond to the 178-kDa CelJ; however, the emPAI results indicate that the CelJ abundance is highest in the cellobiose sample, for which the band appears the faintest on the gel.

The number of unique peptides observed in general increased on cellobiose whereas the contrary was observed on Avicel (Table 11); however, it turns out that the Avicel data set is unreliable. Both the emPAI and RelEx results, inasmuch as the latter can be counted on (see below), show that CelS abundance was lowest in the Avicel sample, even when compared with the cellobiose sample. This contradicts the findings of our previous experiment and the literature, and indicates that the Avicel sample is somehow corrupted. The data from the Avicel sample must therefore be disregarded, along with Avicel versus cellobiose and Avicel versus Avicel with XPM comparisons.

Exoglucanase abundances were highest on Avicel with XPM (about 2 times higher for CelK, and 2.5 times for CelS). CelO was detected on Avicel with XPM only. On the other hand, endoglucanase abundances (except for CelQ) were generally higher on cellobiose and cellobiose with XPM than they were on Avicel with XPM (about 3

times higher for CelA, and 8 times for CelB). The cellobiose with XPM sample, as compared to the cellobiose sample, demonstrated increased expression of CelA (about 1.5-2 times), CelN (6 times), and CelR (2.5 times). Expression of one uncharacterized GH5 (gi 125973339) increased with the addition of XPM, but was highest on cellobiose with XPM (4 times higher than on Avicel with XPM, on which it was 3 times higher than on cellobiose). This GH5 is the same protein that was found in the top-5 emPAI-ranked proteins from the previous experiment.

It is not obvious how or if the addition of XPM affected expression of hemicellulases. Xylanases had the lowest expression on Avicel with XPM. XynA and XynC expression were highest on cellobiose with XPM, whereas XynZ and XynD were highest on cellobiose. As for the other hemicellulases, XghA abundance was highest on cellobiose. A new GH30 had the highest expression on cellobiose with XPM. The greatest number of hemicellulases, including xylanases, was detected on cellobiose (17), followed by cellobiose with XPM (13) and Avicel with XPM (12).

The RelEx data (Table 10) must be discounted due to large standard errors resulting from low numbers of sample to reference peptide ratios calculated. RelEx analysis depends heavily on the number of peptide ratios calculated for statistical significance. The standard errors are so large (50-150%) in this case that it is impossible to distinguish between two sets of sample to reference ratios for the comparison of any protein between any two samples. The numbers of ratios obtained are much lower than in the previous Avicel versus cellobiose experiment, using the same regression and S/N data filters (Table 11). One reason for this, as mentioned above, is the different chromatographic conditions used. Another scenario that might explain the large standard

Table 10. Comparison of relative differences in cellulosome component abundances between two samples, normalized to CipA, as determined by RelEx analysis, from cells grown on Avicel (A) and cellobiose (C) with or without xylan (X), pectin (P), and locust bean gum (M)

		A / C		CXPM / C		AXPM / A		AXPM / CXPM	
GenInfo ID	Protein name & (putative) function/fold(s)	ratio	SE	ratio	SE	ratio	SE	Ratio	SE
125973254	cell-surface anchor	0.95	0.53	0.40	0.36				
125973822	SdbA cell-surface anchor			0.66	0.58				
125975556	CipA scaffoldin	1.00	0.70	1.00	1.23	1.00	0.59	1.00	1.17
125975557	OlpB cell-surface anchor	2.12	1.50	0.75	0.79	0.70	0.67	1.98	2.43
125975558	Orf2p cell-surface anchor	1.04	0.78	0.36	0.34				
125972933	CelK cellobiohydrolase (GH9)	1.17	1.16	1.05	1.46	0.85	0.63	0.94	1.14
125972934	CbhA cellobiohydrolase (GH9)	0.57	0.35	0.82	0.88	1.14	0.59	0.80	0.82
125974579	CelS cellobiohydrolase (GH48)	0.32	0.44	0.53	0.49	4.98	6.69	2.98	2.67
125973055	CelB endoglucanase (GH5)	0.23	0.12	0.86	0.79				
125973315	CelE endoglucanase (GH5), CE2	1.00	0.66	0.49	0.45	1.21	0.67	2.48	2.10
125975353	CelG endoglucanase (GH5)			0.59	0.53				
125972791	CelA endoglucanase (GH8)	0.31	0.16	0.47	0.43	0.87	0.37	0.57	0.48
125972567	CelN endoglucanase (GH9)	0.27	0.14	3.47	3.12	3.27	1.36	0.25	0.21
125973062	CelF endoglucanase (GH9)	0.83	0.47	0.88	0.81				
125973097	CelR endoglucanase (GH9)	1.22	0.78	0.85	0.78	1.34	0.75	1.92	1.65
125973142	CelJ endoglucanase (GH9)	2.91	1.59	0.45	0.40	0.59	0.27	3.83	3.27
125973343	CelD endoglucanase (GH9)			0.96	0.84				
125975294	CelT endoglucanase (GH9)	0.21	0.12	1.16	1.07	4.10	1.71	0.75	0.63
125973339	GH5	2.49	2.89	10.37	9.57	3.44	3.83	0.83	0.71
125974678	GH5			0.20	0.18			1.10	0.91
125972796	GH9	0.25	0.13						
125973263	GH9	0.75	0.38	1.31	1.16	0.73	0.33	0.42	0.36
125975243	GH9	0.50	0.25	0.93	0.81				
125974342	XynC xylanase (GH10)	0.58	0.49	1.14	1.04	0.82	0.64	0.41	0.35
125974464	XynZ xylanase (GH10), CE1	0.81	0.60	1.15	1.23	0.22	0.14	0.15	0.15
125975073	XynD xylanase (GH10)			1.21	1.67			0.36	0.48
125975452	XynA xylanase (GH11), CE4	0.72	0.54	1.09	1.04	0.52	0.31	0.34	0.29
125975293	ManA mannanase (GH26)	0.93	0.47	1.14	1.02	2.14	0.90	1.74	1.47
125975491	GH30			1.12	1.01			0.33	0.28
125973786	GH43	0.14	0.09	1.12	1.16	2.45	1.30	0.30	0.28
125973912	XghA xyloglucanase (GH74)	0.27	0.14	0.48	0.43	0.58	0.25	0.32	0.27

Table 11. Comparison of observed peptide numbers between experiments described in sections 4.2 and 4.3

Sample proteins	Observed/Observable Peptides				N peptide ratios used by RelEx	
	Avicel		cellobiose		previously/here	
	previously	here	previously	here	Avicel	Cellobiose
CipA	42/50	30/34	25/50	28/34	179/29	108/46
CelS (GH48)	29/44	16/23	4/44	21/23	81/3	4/18
CelA (GH8)	14/26	15/13	9/26	17/13	21/8	9/8
CelR (GH9)	28/45	21/26	8/45	19/26	27/15	7/10
CelE (GH5)	24/47	15/36	14/47	14/36	23/6	9/6
GH5 (gi|125973339|)	15/27	6/18	9/27	10/18	14/3	10/7
XynC (GH10)	18/44	16/27	16/44	29/27	21/13	13/24
XynZ (GH10, CE1)	18/44	27/34	25/44	31/34	32/11	48/15

errors is the possibility that the reference culture protein and the sample culture protein (from each of the 4 samples) were degraded to unequal extent. The variance manifests itself here as a drifting average sample to reference ratio within a gel lane. Consider, for example, that if the reference protein is degraded more than the sample protein, the sample to reference ratios calculated would be higher in higher molecular weight gel bands, and lower in lower molecular weight gel bands; this owing to the fact that intact reference protein would be less abundant but degraded protein more abundant. In an effort to diminish the SD, we might consider using only ratios from gel bands with molecular weight equal to and greater than that of a given protein to calculate its average sample to reference ratio, thus incorporating only so-called intact protein in the final analysis. However, in so doing we would be biasing against the true protein abundance, to which the degraded protein contributes; still, if we assume that all the sample protein is degraded (or not) to the same extent, this should not be problematic since the same amount of reference protein, degraded or not, was mixed with all samples. This data manipulation was attempted, but the already low number of peptide ratios made it unfruitful. Ultimately, my sense is that, even though these data cannot be used because of the staggeringly large standard errors, the ratios are most likely correct because it agrees roughly with the emPAI data.

4.3.2. Enzymatic activities of cellulosomes isolated from cultures grown with xylan, pectin, and locust bean gum

Cellulosomes were also isolated from each of the 4 unmixed sample cultures in order to evaluate differences in their specific exoglucanase, endoglucanase and xylanase

activities (Table 12). All activities were measured at pH 5.7 and 60°C. Even though it is not assumed that all enzymes of a given specificity have the same activity, relative differences in enzymatic activities were expected to agree more or less with relative differences in enzyme abundances as per the protein-specific MS data (section 4.3.1). As mentioned above, the Avicel sample should be disregarded based on the MS results which showed, contrary to the literature and our earlier findings, that CelS levels were lower in the Avicel sample than in the cellobiose sample.

Exoglucanase-like activities, which were measured using the chromogenic cellobiose derivative PNPC, did not change from sample to sample (Table 12). The PNPCase activities observed are up to 5-fold higher than the value of 0.46 U/mg reported for a "subcellulosome fraction" comprising 6 main subunits [100]. Activity against PNPC has been shown for individual components; for example, PNPCase of 15.1 U/mg was reported for recombinant CelK at pH 6.0 and 65°C [98]. Cellulosomal specific activities are lower in part because there are only 4 known exoglucanases out of more than 20 known docking enzymes in the very large cellulosome complex (Table 9). Also, it has previously been shown that CelS, a cellobiohydrolase and the cellulosome's most important exoglucanase, has no activity against PNPC [165] and cannot degrade anything smaller than a cellotetraose [97].

Endoglucanase activity was measured using CMC, which exoglucanases typically cannot break down. The specific CMCase activity of cellobiose-grown cellulosomes was 42 U/mg, about double that for any other sample (Table 12). This value falls within the typical range previously observed for cellulosome preparations [166, 167]. The MS data above did not suggest a dramatic difference between the overall endoglucanase activities

Table 12. Specific exoglucanase, endoglucanase, and xylanase activities at pH 5.7 and 60°C for cellulosomes grown on Avicel (A) and cellobiose (C) with or without xylan (X), pectin (P), and locust bean gum (M)

		Specific activity ± SE (μmol/min/mg)		
Growth Substrate	Total protein ± SD (μg/mL)	Exoglucanase (PNPCase)	Endoglucanase (CMCase)	Xylanase
C	1755 ± 118	2.18 ± 0.29	41.84 ± 5.06	13.27 ± 1.55
CXPM	644 ± 29	1.15 ± 0.09	19.75 ± 1.52	11.54 ± 0.89
A	723 ± 175	1.84 ± 0.91	18.60 ± 7.85	12.88 ± 5.52
AXPM	3806 ± 579	2.17 ± 0.67	17.20 ± 4.66	6.44 ± 1.70

of any of the samples (Table 9). If any sample might have been expected to demonstrate higher activity based on the MS results, it was the cellobiose with XPM sample. In spite of the changing enzyme abundances between samples, overall activity was expected to remain more or less constant between all samples, as had been observed previously [30, 46]. It is possible that the elevated activity of the cellobiose sample was due to a highly active subunit present in greater abundance in that sample, namely CelD, CelG or CelJ (Table 9), or some combination thereof.

Cellulosomal xylanase activity was tested using xylan from birch wood. A previous study reported a value of 100 U/mg for cellulosomes purified from cellobiose- or cellulose-grown cultures [31]. The reported value is almost one order of magnitude greater than the values observed here (Table 12); however, a different strain (YS rather than ATCC 27405) and a chromogenic rather than a reducing sugar assay (using the xylan derivative Remazol Brilliant Blue R-D-xylan) were used at pH 6-7.5 and 70°C. Cellulosomes grown on Avicel with XPM displayed the lowest overall xylanase activity. While this result corroborates the MS data described above (Table 9), it is not clear whether or not the presence of XPM acted to repress xylanase expression and whether or not the presence of cellobiose, in the case of the cellobiose with XPM sample, helped to counteract this effect.

5. DISCUSSION

This thesis presents the most comprehensive proteomic study of the *C. thermocellum* cellulosome to date. Until the recent use of two-dimensional gels and MS-based methods to improve the compositional detail of the *C. thermocellum* cellulosome [36, 91], most of the work concerning the identification of cellulosomal components had so far been done by means of enzymatic assay [68] or Western blot analysis [99, 108-110, 152, 153, 155-159, 161, 162]. The detection of 29 (16 in section 3.2, 13 more in section 3.1) new Doc1-containing proteins represents a 130 percent increase in the number of docking subunits observed in cellulosomes. However, it should be noted that in general the proteins detected in highest abundance were known, which attests to the fact that the more abundant proteins are the more 'discoverable'. Yet one new GH5 enzyme (gi 125973339) containing a predicted galactose-binding domain was found in considerable abundance under both growth conditions, and may prove to be a subunit of some importance upon further investigation.

The first part of this discussion focuses on the comparison of cellulosomes from Avicel- and cellobiose-grown cells described in section 3.2. The three known docking subunits to escape detection were the non-catalytic docking component CseP [110], the serine protease inhibitor PinA [109], and the bifunctional component CelH [36]; however, all three of these were observed by us in earlier trials (data not shown) in which either no reference protein was mixed in or the reference had not been ^{15}N-enriched to 99%. CseP and PinA were detected on both substrates, whereas CelH, which has both a GH5 and a GH26 domain, was detected only on cellobiose. CelO, the only known GH5

exoglucanase in *C. thermocellum* [105], is the only previously cloned docking gene product never to be detected by us.

XynD was detected exclusively on cellobiose even though it had been discovered on cellulose by 2D gel followed by MS [91]; and ManA and LicB were detected exclusively on Avicel, whereas they had previously been observed on cellobiose by Western blot analysis [158, 159]. These discrepancies could be explained by the differences between the protein identification methods used in the previous studies and that used in the present work. In Western blot analysis, a radiolabeled antibody raised against a particular subunit can be used to detect that subunit, even in low concentrations, directly on an SDS-PAGE gel, despite the presence of other proteins. In LC-MS, on the other hand, the high specificity only applies at the database screening stage, while the detection of a protein depends on several considerations including its relative abundance, the efficiency of its proteolysis, and, in our case, the extraction efficiency of its derivative peptides from the acrylamide gel. A peptide present at a low but in theory detectable concentration may not be detected if a more abundant peptide elutes from the LC column at the same time. Compounding these factors, the presence of ^{15}N-labelled peptides in our experimental setup in fact doubled the complexity of each sample; for even even though they did not count in the identification of cellulosomal proteins, they were detected by the MS. It is possible that XynD, ManA and LicB were present in both samples but that their peptide signals were masked by peptides from proteins present in higher abundance.

Growth on the different substrates revealed a similar mix of cellulosomal components that were present in significantly different relative amounts. Differences in the relative expression levels of individual components grown on either carbon source

demonstrated GH family-specific regulatory patterns, providing evidence in support of existing hypotheses for cellulosomal component regulation as well as contributing a novel distinction with respect to endoglucanase synthesis.

The exoglucanase CelS exhibited the greatest increase of any docking component during growth on Avicel as compared to cellobiose. The increase of CelS on Avicel versus cellobiose had already been observed at the protein level by SDS-PAGE [84] and Western blot analysis [1]. This result also agrees with changes in *celS* transcript levels per cell between growth on cellulose and cellobiose [1]. Exoglucanases are the key enzymes in cellulase mixtures effective on crystalline cellulose [2], so it was not surprising that exoglucanase CelK also increased on Avicel, even while the expression of CbhA did not change significantly.

Docking proteins with known endoglucanase activity demonstrated varied expression patterns. GH5 endoglucanases CelB, CelE and CelG demonstrated higher expression when cells were grown on cellobiose than on Avicel. The same was true for CelA from GH8. In contrast, CelJ from GH9 showed increased expression on Avicel, while that of other GH9 endoglucanases CelF, CelR and CelT did not change significantly. The detection of CelN and CelQ on Avicel and not cellobiose may be taken as another indication of increased GH9 endoglucanase production on Avicel. The differential expression of GH9 versus GH5 endoglucanases poses an apparent discrepancy with the recent transcript analysis of Dror *et al.* [120], who observed increased transcript levels per cell of each of the endoglucanase genes *celB* and *celG* from GH5 and *celD* from GH9 when cells were grown at low versus high growth rate and also on cellulose versus cellobiose. Thus, while our results with respect to GH9

endoglucanases agree with these previous findings at the transcript level, the increase of GH5 endoglucanases and of CelA on cellobiose was a somewhat unanticipated result. One possible explanation for the difference between the trends observed at the mRNA and protein levels is that GH9 endoglucanase genes may be more responsive to catabolite repression than *celA* or GH5 endoglucanase genes, such that the former would be more repressed on cellobiose than either of the latter.

The data suggest the organism has a "cellulolytic preference" for GH9 endoglucanases when degradation of crystalline cellulose is required. In total, cellulosomal GH9 cellulases contained in the *C. thermocellum* genome outnumber GH5 enzymes by 14-to-8. This preference could be due to what distinguishes them from CelA and GH5 endoglucanases: the presence, in many instances, of a type IIIc carbohydrate binding module, which has been shown to participate in the catalytic activity of the enzyme [107, 108] and be responsible for processivity [65, 74]. What is more, GH9 endoglucanases carry out different modes of attack on cellulose, resulting in cellodextrins of different lengths [107]. CelR, which was the most abundant endoglucanase in cellulosomes from Avicel-grown cells, is one such enzyme, a processive GH9 endoglucanase that produces cellotetraose as its primary hydrolysis product [94], which is more energetically favourable for the cell than production of cellobiose [62].

Thus, the apparent constitutive nature of overall endoglucanase activity appears to be the result of different GH5, GH8 or GH9 endoglucanases varying in expression to balance out global activity. Still, repression of exoglucanase expression and activity by cellobiose holds. It is possible that the differences observed between the two samples were diminished by an evening out effect proportional to the titering of cellobiose outside

the cell. In the cellulose-grown culture, slow growth or lack of a repressor molecule (cellobiose) initially lead to up-regulation of genes for activity against crystalline cellulose. As the Avicel was degraded, the cellobiose concentration accumulated, repressing these genes. Equally, in the cellobiose-grown culture, it is possible that as the cellobiose concentration became limited, genes for enzymes crucial to crystalline cellulose degradation were activated.

With respect to hemicellulases, all subunits with xylanase or xyloglucanase activity decreased on Avicel, as per RelEx and emPAI. XynC production has previously been shown to increase on cellobiose [84, 120], and *xynC* transcript levels have been found to increase on cellobiose in a growth rate-independent fashion [120]. In this study, XynZ, XynA, XynC and XghA were among the five most abundant docking components in cellobiose-grown cellulosomes, along with CelA. XynD and XynY were not detected in the Avicel sample, possibly because their signals were overwhelmed by those of more abundant subunits. On the other hand, their exclusive detection on cellobiose might be taken as another indication of increased xylanase production on cellobiose. Other new subunits with glycosidase and arabinofuranosidase activities were detected exclusively on cellobiose. The trend of increased hemicellulase production on cellobiose could also explain the increase of bifunctional subunit CelE, which has a family 2 carbohydrate esterase domain in addition to a GH5. As for other hemicellulases, no change was noted for ChiA, and the appearance of LicB and ManA on Avicel but not cellobiose suggests that transcription of these genes was repressed on cellobiose. In the case of *manA*, Stevenson *et al.* [119] reported a 10-fold reduction in its transcript level on cellobiose as compared to cellulose. Thus, while xylanase transcription is growth-rate independent and

increases on cellobiose, chitinase, lichenase and mannanase appear to be under a different type of regulation mechanism. *C. thermocellum* is unable to utilize the pentose sugars produced by the action of xylanases and other hemicellulases [26, 30]; therefore, the apparent role of hemicellulases is to expose cellulose to the action of cellulases. When the organism is not mining energy from cellulose, as when it is grown on cellobiose, in general it appears to prepare itself to mine cellulose from plant wall materials, hemicellulose and lignin, as it would in its natural ecosystem.

Finally, our investigation into the addition of xylan, pectin and locust bean gum (galactomannan) to Avicel- and cellobiose-grown cultures proved inconclusive (section 3.3). It was not clear what effect, if any, the addition of these hemicelluloses has on the cellulosomal subunit profile. The expectation was that xylanase, pectinase, and mannanase expression would increase. While an additional 12 components were detected, including a new GH30 subunit, 2 new lipases and a pectate lyase, these were not exclusively in the samples containing the hemicelluloses. The only hint of a regulating quality their presence may have effected was that they appeared to repress xylanase expression. This possibility runs contrary to what is known about hemicellulolytic enzyme production. Analogous to cellulase induction of endoglucanases by cellobiose, hemicellulases are thought to be induced by the presence of low levels of their end-products, which can enter the cell and stimulate their transcription [168]. This is the case in xylanolytic xylose-utilizing organisms [169-171] and for *C. thermocellum*'s close neighbour *Clostridium cellulovorans* [122, 123]. However, it may be that since *C. thermocellum* cannot utilize xylose or xylobiose as carbon source, it does not possess the machinery for control of xylanases by xylan metabolites.

In conclusion, this work provides a global view of the *C. thermocellum* cellulosome. During growth on two utilizable carbon sources with and without hemicelluloses added, the organism produced a wide variety of dockable hydrolytic enzymes, accounting for more than 80% of the genes containing dockerin I sequences. Of the remaining unobserved putative dockable gene products, there remain about 12 proteins of unknown function, which may be inducible using more complex substrates. Future work for this project should begin with uncovering the quantitative differences in the cell surface, cell membrane and cytosolic protein fractions of *C. thermocellum*, grown on Avicel or cellobiose; these subproteomes should be obtained with no difficulty (Figure 9) and should reveal more detail as to cellular mechanisms underlying cellulosome regulation and metabolism of the products of its action on various substrates. An understanding of the mechanisms by which bacteria regulate the expression of the various cellulases and hemicellulases at their disposal will be important to the eventual production of optimal enzyme cocktails or designer cellulosomes used in the break-down of cellulosic materials for the transition from an oil-based to a carbohydrate-based economy.

6. REFERENCES

1. Dror, T.W., et al., *Regulation of the cellulosomal celS (cel48A) gene of Clostridium thermocellum is growth rate dependent.* J. Bacteriol., 2003. **185**(10): p. 3042-3048.
2. Teeri, T.T., *Crystalline cellulose degradation: new insight into the function of cellobiohydrolases.* Trends Biotechnol., 1997. **15**(5): p. 160-167.
3. Gold, N.D. and V.J.J. Martin, *Global view of the Clostridium thermocellum cellulosome revealed by quantitative proteomic analysis.* J. Bacteriol., 2007. **189**(19): p. 6787-6795.
4. Watson, R.T., et al., *Climate Change 2001: Synthesis Report.* 2001, Intergovernmental Panel on Climate Change: Geneva, Switzerland. p. 184.
5. Sorkin, L., *Climate Change Impacts Rise*, in *Vital Signs 2006-2007.* 2006, Worldwatch Institute. p. 42-43.
6. Canada, Environment Canada, *Canada's 2005 National Greenhouse Gas Inventory: A Summary of Trends (1990-2005).* 2007. Available from: http://www.ec.gc.ca/pdb/ghg/inventory_report/2005/2005summary_e.pdf.
7. Canada, Environment Canada, *Canada's Fourth National Report on Climate Change: Actions to Meet Commitments Under the United Nations Framework Convention on Climate Change.* 2006. Available from: http://www.ec.gc.ca/climate/4th_Report_on_CC_e.pdf. p. 318.
8. Yuksel, F. and B. Yuksel, *The use of ethanol-gasoline blend as a fuel in an SI engine.* Renew. Energ., 2004. **29**(7): p. 1181-1191.

9. Dias De Oliveira, M.E., B.E. Vaughan, and E.J. Rykiel, *Ethanol as fuel: energy, carbon dioxide balances, and ecological footprint.* BioScience, 2005. **55**(7): p. 593-602.

10. Farrell, A.E., et al., *Ethanol can contribute to energy and environmental goals.* Science, 2006. **311**(5760): p. 506-508.

11. Wyman, C.E. and N.D. Hinman, *Ethanol. Fundamentals of production from renewable feedstocks and use as a transportation fuel.* Appl. Biochem. Biotech., 1990. **24-25**: p. 735-753.

12. Lynd, L.R., et al., *Microbial cellulose utilization: fundamentals and biotechnology.* Microbiol. Mol. Biol. R., 2002. **66**(3): p. 506-577.

13. Lynd, L.R., et al., *Fuel ethanol from cellulosic biomass.* Science, 1991. **251**(4999): p. 1318-1323.

14. Champagne, P., *Feasibility of producing bio-ethanol from waste residues: a Canadian perspective.* Resour. Conserv. Recy., 2007. **50**: p. 211-230.

15. SunOpta BioProcess Inc., *Presentation at EPAC Conference*. June 11, 2007 [cited October 31, 2007]; Available from: http://www.sunopta.com/uploadedFiles/bio-process/News_and_Events/070611%20EPAC%20Conference.pdf.

16. Abengoa Bioenergy, *Abengoa Bioenergy opens pilot plant for the energy of the future.* October 2007 [cited October 31, 2007]; Available from: http://www.abengoabioenergy.com/about/index.cfm?page=15&lang=1&headline=60.

17. Iogen Corporation. *About Iogen.* 2007 [cited October 31, 2007]; Available from: http://www.iogen.ca/company/about/index.html.

18. Mascoma Corporation. *Mascoma Corporation to build nation's first switchgrass cellulosic ethanol plant*. September 27, 2007 [cited October 31, 2007]; Available from: http://www.mascoma.com/welcome/pdf/09.27.07%20-%20Mascoma%20News%20Release%20-TENN%20-%20FiNAL.pdf.

19. Saha, B.C., *Hemicellulose bioconversion.* J. Ind. Microbiol. Biotechnol., 2003. **30**: p. 279-291.

20. Grabber, J.H., *How do lignin composition, structure, and cross-linking affect degradability? A review of cell wall model studies.* Crop. Sci., 2005. **45**(3): p. 820-831.

21. Shallom, D. and Y. Shoham, *Microbial hemicellulases.* Curr. Opin. Microbiol., 2003. **6**: p. 219-228.

22. Najafpour, G. and H. Younesi, *Ethanol and acetate synthesis from waste gas using batch culture of Clostridium ljungdahlii.* Enzyme Microb. Tech., 2006. **38**(1-2): p. 223-228.

23. Palmqvist, E. and B. Hahn-Hagerdal, *Fermentation of lignocellulosic hydrolysates. I: inhibition and detoxification.* Bioresource Technol., 2000. **74**(1): p. 17-24.

24. Lynd, L.R., *Overview and evaluation of fuel ethanol from cellulosic biomass: technology, economics, the environment, and policy.* Annu. Rev. Energy Env., 1996. **21**(1): p. 403-465.

25. Ethanol Producer Magazine, *Enzyme contract concludes successfully*. June 2005. [cited October 31, 2007]; Available from:

http://ethanolproducer.com/article.jsp?article_id=201&q=novozymes%20genencor&category_id=37.

26. Demain, A.L., M. Newcomb, and J.H.D. Wu, *Cellulase, clostridia, and ethanol.* Microbiol. Mol. Biol. R., 2005. **69**(1): p. 124-154.

27. Lynd, L.R., et al., *Consolidated bioprocessing of cellulosic biomass: an update.* Curr. Opin. Biotech., 2005. **16**: p. 577-583.

28. Lynd, L.R. and H.G. Grethlein., *Hydrolysis of dilute acid pretreated hardwood and purified microcrystalline cellulose by cell-free broth from Clostridium thermocellum.* Biotechnol. Bioeng. Symp., 1987. **29**: p. 92-100.

29. Lamed, R. and J.G. Zeikus, *Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in Clostridium thermocellum and Thermoanaerobium brockii.* J. Bacteriol., 1980. **144**: p. 569-578.

30. Garcia-Martinez, D.V., et al., *Studies on cellulase production by Clostridium thermocellum.* Eur. J. Appl. Microbiol. Biot., 1980. **9**: p. 189-197.

31. Morag, E., E.A. Bayer, and R. Lamed, *Relationship of cellulosomal and noncellulosomal xylanases of Clostridium thermocellum to cellulose-degrading enzymes.* J. Bacteriol., 1990. **172**(10): p. 6098-6105.

32. Venkateswaren, S. and A.L. Demain, *The Clostridium thermocellum-Clostridium thermosaccharolyticum ethanol production process: nutritional studies and scale-down.* Chem. Eng. Commun., 1986. **45**: p. 53-60.

33. Ng, T.K., A. Ben-Bassat, and J.G. Zeikus., *Ethanol production by thermophilic bacteria: fermentation of cellulosic substrates by cocultures of Clostridium*

thermocellum and Clostridium thermohydrosulfuricum. Appl. Environ. Microbiol., 1981. **41**: p. 1337-1343.

34. Tyurin, M.V., S.G. Desai, and L.R. Lynd, *Electrotransformation of Clostridium thermocellum.* Appl. Environ. Microb., 2004. **70**(2): p. 883-890.

35. Heap, J.T., et al., *The ClosTron: a universal gene knock-out system for the genus Clostridium.* J. Microbiol. Meth., 2007. **70**(3): p. 452-464.

36. Williams, T.I., et al., *Proteomic profile changes in membranes of ethanol-tolerant Clostridium thermocellum.* Appl. Microbiol. Biot., 2007. **V74**(2): p. 422-432.

37. Brown, S., et al., *Construction and evaluation of a Clostridium thermocellum ATCC 27405 whole-genome oligonucleotide microarray.* Appl. Biochem. Biotech., 2007. **137-140**(1): p. 663-674.

38. Shoseyov, O., et al., *Primary sequence analysis of Clostridium cellulovorans cellulose binding protein A.* Proc. Natl. Acad. Sci. USA, 1992. **89**: p. 3483-3487.

39. Pages, S., et al., *Sequence analysis of scaffolding protein CipC and ORFXp, a new cohesin-containing protein in Clostridium cellulolyticum: comparison of various cohesin domains and subcellular localization of ORFXp.* J. Bacteriol., 1999. **181**(6): p. 1801-1810.

40. Kakiuchi, M., et al., *Cloning and DNA sequencing of the genes encoding Clostridium josui scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome.* J. Bacteriol., 1998. **180**: p. 4303-4308.

41. Sabathé, F., A. Belaich, and P. Soucaille, *Characterization of the cellulolytic complex (cellulosome) of Clostridium acetobutylicum.* FEMS Microbiol. Lett., 2002. **217**(1): p. 15-22.

42. Xu, Q., et al., *The cellulosome system of Acetivibrio cellulolyticus includes a novel type of adaptor protein and a cell surface anchoring protein.* J. Bacteriol., 2003. **185**(15): p. 4548-4557.

43. Devillard, E., et al., *Ruminococcus albus 8 mutants defective in cellulose degradation are deficient in two processive endocellulases, Cel48A and Cel9B, both of which possess a novel modular architecture.* J. Bacteriol., 2004. **186**(1): p. 136-145.

44. Rincon, M.T., et al., *Unconventional mode of attachment of the Ruminococcus flavefaciens cellulosome to the cell surface.* J. Bacteriol., 2005. **187**(22): p. 7569-7578.

45. Shoham, Y., R. Lamed, and E.A. Bayer, *The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides.* Trends in Microbiology, 1999. **7**(7): p. 275-281.

46. Shinmyo, A., D.V. Garcia-Martinez, and A.L. Demain, *Studies on the extracellular cellulolytic enzyme complex produced by Clostridium thermocellum.* J. Appl. Biochem., 1979. **1**: p. 202-209.

47. Murashima, K., A. Kosugi, and R.H. Doi, *Synergistic effects on crystalline cellulose degradation between cellulosomal cellulases from Clostridium cellulovorans.* J. Bacteriol., 2002. **184**(18): p. 5088-5095.

48. Murashima, K., A. Kosugi, and R.H. Doi, *Synergistic effects of cellulosomal xylanase and cellulases from Clostridium cellulovorans on plant cell wall degradation.* J. Bacteriol., 2003. **185**(5): p. 1518-1524.

49. Fierobe, H.P., et al., *Degradation of cellulose substrates by cellulosome chimeras. Substrate targeting versus proximity of enzyme components.* J. Biol. Chem., 2002. **277**: p. 49621–30.

50. Fierobe, H.-P., et al., *Action of designer cellulosomes on homogeneous versus complex substrates: controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin.* J. Biol. Chem., 2005. **280**(16): p. 16325-16334.

51. Hammel, M., et al., *Structural basis of cellulosome efficiency explored by small angle X-ray scattering.* J. Biol. Chem., 2005. **280**(46): p. 38562-38568.

52. Kleman-Leyer, K.M., et al., *The cellulases endoglucanase I and cellobiohydrolase II of Trichoderma reesei act synergistically to solubilize native cotton cellulose but not to decrease its molecular size.* Appl. Environ. Microbiol., 1996. **62**(8): p. 2883-2887.

53. Mingardon, F., et al., *Incorporation of fungal cellulases in bacterial minicellulosomes yields viable, synergistically acting cellulolytic complexes.* Appl. Environ. Microbiol., 2007. **73**(12): p. 3822-3832.

54. Strobel, H.J., F.C. Caldwell, and K.A. Dawson, *Carbohydrate transport by the anaerobic thermophile Clostridium thermocellum LQRI.* Appl. Environ. Microb., 1995. **61**(11): p. 4012-4015.

55. Lu, Y., Y.-H.P. Zhang, and L.R. Lynd, *Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum.* P. Natl. Acad. Sci. USA, 2006. **103**(44): p. 16165-16169.

56. Kruus, K., et al., *Product inhibition of the recombinant CelS, an exoglucanase component of the Clostridium thermocellum cellulosome.* Appl. Microbiol. Biot., 1995. **44**(3): p. 399-404.

57. Grabnitz, F. and W.L. Staudenbauer, *Characterization of two β-glucosidase genes from Clostridium thermocellum.* Biotechnol. Lett., 1988. **10**: p. 73-78.

58. Sheth, K. and J.K. Alexander, *Purification and properties of β-1,4-oligoglucan:orthophosphate glucosyltransferase from Clostridium thermocellum.* J. Biol. Chem., 1969. **244**: p. 457-464.

59. Alexander, J.K., *Purification and specificity of cellobiose phosphorylase from Clostridium thermocellum.* J. Biol. Chem., 1968. **243**: p. 2899-2904.

60. Nochur, S.V., et al., *Mode of sugar phosphorylation in Clostridium thermocellum.* Appl. Biochem. Biotech., 1992. **33**: p. 33-41.

61. Zhang, Y.-H.P. and L.R. Lynd, *Kinetics and relative importance of phosphorolytic and hydrolytic cleavage of cellodextrins and cellobiose in cell extracts of Clostridium thermocellum.* Appl. Environ. Microb., 2004. **70**(3): p. 1563-1569.

62. Zhang, Y.-H.P. and L.R. Lynd, *Cellulose utilization by Clostridium thermocellum: bioenergetics and hydrolysis product assimilation.* P. Natl. Acad. Sci. USA, 2005. **102**(20): p. 7321-7325.

63. Coughlan, M.P., et al., *The cellulolytic enzyme complex of is very large.* Biochem. Bioph. Res. Co., 1985. **130**(2): p. 904-909.

64. Mayer, F., et al., *Macromolecular organization of the cellulolytic enzyme complex of Clostridium thermocellum as revealed by electron microscopy.* Appl. Environ. Microb., 1987. **53**(12): p. 2785-2792.

65. Bayer, E.A., et al., *Cellulosomes - structure and ultrastructure.* J. Struct. Biol., 1998. **124**(2-3): p. 221-234.

66. Johnson, E.A. and A.L. Demain, *Probable involvement of sulfhydryl groups and a metal as essential components of the cellulase of Clostridium thermocellum.* Arch. Microbiol., 1984. **137**: p. 135-138.

67. Johnson, E.A., et al., *Saccharification of complex cellulosic substrates by the cellulase system from Clostridium thermocellum.* Appl. Environ. Microb., 1982. **43**(5): p. 1125-1132.

68. Wu, J.H.D., W.H. Orme-Johnson, and A.L. Demain, *Two components of an extracellular protein aggregate of Clostridium thermocellum together degrade crystalline cellulose.* Biochemistry-US, 1988. **27**(5): p. 1703-1709.

69. Bayer, E.A., E. Morag, and R. Lamed, *The cellulosome - a treasure-trove for biotechnology.* Trends Biotechnol., 1994. **12**(9): p. 379-386.

70. Gerngross, U.T., et al., *Sequencing of a Clostridium thermocellum gene (cipA) encoding the cellulosomal SL-protein reveals an unusual degree of internal homology.* Mol. Microbiol., 1993. **8**(2): p. 325-334.

71. Gerwig, G.J., et al., *Novel O-linked carbohydrate chains in the cellulase complex (cellulosome) of Clostridium thermocellum. 3-O-Methyl-N-acetylglucosamine as a constituent of a glycoprotein.* J. Biol. Chem., 1989. **264**: p. 1027-1035.

72. Gerwig, G.J., et al., *The nature of the carbohydrate-peptide linkage region in glycoproteins from the cellulosomes of Clostridium thermocellum and Bacteroides cellulosolvens.* J. Biol. Chem., 1993. **268**: p. 26956-26960.

73. Kruus, K., et al., *The anchorage function of CipA (CelL), a scaffolding protein of the Clostridium thermocellum cellulosome.* P. Natl. Acad. Sci. USA, 1995. **92**(20): p. 9254-9258.

74. Tormo, J., et al., *Crystal Structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose.* EMBO J., 1996. **15**: p. 5739-5751.

75. Carrard, G., et al., *Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose.* Proc. Natl. Acad. Sci. USA, 2000. **97**: p. 10342-10347.

76. Din, N., et al., *Non-hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulose.* Bio/Technology, 1991. **9**: p. 1096-1099.

77. Pages, S., et al., *Role of scaffolding protein CipC of Clostridium cellulolyticum in cellulose degradation.* J. Bacteriol., 1997. **179**(9): p. 2810-2816.

78. Salamitou, S., et al., *Subcellular localization of Clostridium thermocellum ORF3p, a protein carrying a receptor for the docking sequence borne by the catalytic components of the cellulosome.* J. Bacteriol., 1994. **176**(10): p. 2828-2834.

79. Schaeffer, F., et al., *Duplicated dockerin subdomains of Clostridium thermocellum endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity.* Biochemistry-US, 2002. **41**(7): p. 2106-2114.

80. Mechaly, A., et al., *Cohesin-dockerin interaction in cellulosome assembly. A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition.* J. Biol. Chem., 2001. **276**(13): p. 9883-9888.

81. Adams, J.J., et al., *Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex.* P. Natl. Acad. Sci. USA, 2006. **103**(2): p. 305-310.

82. Jindou, S., et al., *Interaction between a type-II dockerin domain and a type-II cohesin domain from Clostridium thermocellum cellulosome.* Biosci. Biotech. Bioch., 2004. **68**(4): p. 924-926.

83. Janin, J., *Kinetics and thermodynamics of protein-protein interactions*, in *Protein-Protein Recognition*, C. Kleanthous, Editor. 2000, Oxford University Press. p. 1-26.

84. Bayer, E.A., E. Setter, and R. Lamed, *Organization and distribution of the cellulosome in Clostridium thermocellum.* J. Bacteriol., 1985. **163**(2): p. 552-559.

85. Fierobe, H.-P., et al., *Design and production of active cellulosome chimeras: selective incorporation of dockerin-containing enzymes into defined functional complexes.* J. Biol. Chem., 2001. **276**(24): p. 21257-21261.

86. Pagès, S., et al., *Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of*

specificity determinants of the dockerin domain. Proteins Struct. Funct. Genet., 1997. **29**(4): p. 517-527.

87. Shimon, L.J.W., et al., *A cohesin domain from Clostridium thermocellum: the crystal structure provides new insights into cellulosome assembly.* Structure, 1997. **5**(3): p. 381-390.

88. Tavares, G.A., P. Beguin, and P.M. Alzari, *The crystal structure of a type I cohesin domain at 1.7 Å resolution.* J. Mol. Biol., 1997. **273**(3): p. 701-713.

89. Lytle, B.L., et al., *Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain.* J. Mol. Biol., 2001. **307**(3): p. 745-753.

90. Carvalho, A.L., et al., *Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex.* P. Natl. Acad. Sci. USA, 2003. **100**(24): p. 13809-13814.

91. Zverlov, V.V., J. Kellermann, and W.H. Schwarz, *Functional subgenomics of Clostridium thermocellum cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes.* Proteomics, 2005. **5**(14): p. 3646-3653.

92. Henrissat, B. and G. Davies, *Structural and sequence-based classification of glycoside hydrolases.* Curr. Opin. Struc. Biol., 1997. **7**(5): p. 637-644.

93. Coutinho, P.M. and B. Henrissat, *Carbohydrate-active enzymes: an integrated database approach.* Recent Advances in Carbohydrate Bioengineering, ed. H.J. Gilbert, et al. 1999, Cambridge: The Royal Society of Chemistry. 3-12.

94. Zverlov, V.V., N. Schantz, and W.H. Schwarz, *A major new component in the cellulosome of Clostridium thermocellum is a processive endo-β-1,4-glucanase producing cellotetraose.* FEMS Microbiol. Lett., 2005. **249**(2): p. 353-358.

95. Fuchs, K.-P., et al., *Lic16A of Clostridium thermocellum, a non-cellulosomal, highly complex endo-{beta}-1,3-glucanase bound to the outer cell surface.* Microbiology, 2003. **149**(4): p. 1021-1031.

96. Blum, D.L., et al., *Feruloyl esterase activity of the Clostridium thermocellum cellulosome can be attributed to previously unknown domains of XynY and XynZ.* J. Bacteriol., 2000. **182**(5): p. 1346-1351.

97. Morag, E., et al., *Isolation and properties of a major cellobiohydrolase from the cellulosome of Clostridium thermocellum.* J. Bacteriol., 1991. **173**(13): p. 4155-4162.

98. Kataeva, I., et al., *Cloning and sequence analysis of a new cellulase gene encoding CelK, a major cellulosome component of Clostridium thermocellum: evidence for gene duplication and recombination.* J. Bacteriol., 1999. **181**(17): p. 5288-5295.

99. Lemaire, M. and P. Beguin, *Nucleotide sequence of the celG gene of Clostridium thermocellum and characterization of its product, endoglucanase CelG.* J. Bacteriol., 1993. **175**(11): p. 3353-3360.

100. Kobayashi, T., et al., *Subcellulosome preparation with high cellulase activity from Clostridium thermocellum.* Appl. Environ. Microbiol., 1990. **56**(10): p. 3040-3046.

101. Morag, E., et al., *Cellulase S-S (CelS) is synonymous with major cellobiohydrolase (subunit S8) from the cellulosome of Clostridium thermocellum.* Appl. Biochem. Biotech., 1993. **43**: p. 147-151.

102. Wang, W.K. and J.H. Wu, *Structural features of the Clostridium thermocellum cellulase S_S gene.* Appl. Biochem. Biotech., 1993. **39-40**: p. 149-58.

103. Lamed, R., E. Setter, and E.A. Bayer, *Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum.* J. Bacteriol., 1983. **156**: p. 828-836.

104. Guimaraes, B.G., et al., *The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the Clostridium thermocellum cellulosome.* J. Mol. Biol., 2002. **320**: p. 587-596.

105. Zverlov, V.V., G.A. Velikodvorskaya, and W.H. Schwarz, *A newly described cellulosomal cellobiohydrolase, CelO, from Clostridium thermocellum: investigation of the exo-mode of hydrolysis, and binding capacity to crystalline cellulose.* Microbiology, 2002. **148**(1): p. 247-255.

106. Barr, B.K., et al., *Identification of two functionally different classes of exocellulases.* Biochemistry, 1996. **35**: p. 586-592.

107. Arai, T., et al., *Properties of cellulosomal family 9 cellulases from Clostridium cellulovorans.* Appl. Microbiol. Biot., 2006. **71**(5): p. 654-660.

108. Arai, T., et al., *Sequence of celQ and properties of CelQ, a component of the Clostridium thermocellum cellulosome.* Appl. Microbiol. Biot., 2001. **57**(5): p. 660-666.

109. Kang, S., et al., *The functional repertoire of prokaryote cellulosomes includes the serpin superfamily of serine proteinase inhibitors.* Mol. Microbiol., 2006. **60**(6): p. 1344-1354.

110. Zverlov, V.V., G.A. Velikodvorskaya, and W.H. Schwarz, *Two new cellulosome components encoded downstream of celI in the genome of Clostridium thermocellum: the non-processive endoglucanase CelN and the possibly structural protein CseP.* Microbiology, 2003. **149**(2): p. 515-524.

111. Bayer, E.A., et al., *The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides.* Annu. Rev. Microbiol., 2004. **58**(1): p. 521-554.

112. Zhang, Y.-H.P. and L.R. Lynd, *Regulation of cellulase synthesis in batch and continuous cultures of Clostridium thermocellum.* J. Bacteriol., 2005. **187**(1): p. 99-106.

113. Lamed, R., et al., *Major characteristics of the cellulolytic system of Clostridium thermocellum coincide with those of the purified cellulosome.* Enzyme Microb. Tech., 1985. 7(1): p. 37-41.

114. Freier, D., C.P. Mothershed, and J. Wiegel, *Characterization of Clostridium thermocellum JW20.* Appl. Environ. Microb., 1988. **54**(1): p. 204-211.

115. Mishra, S., P. Beguin, and J.P. Aubert, *Transcription of Clostridium thermocellum endoglucanase genes celF and celD.* J. Bacteriol., 1991. **173**(1): p. 80-85.

116. Warner, J.B. and J.S. Lolkema, *CcpA-dependent carbon catabolite repression in bacteria.* Microbiol. Mol. Biol. R., 2003. **67**(4): p. 475-490.

117. Johnson, E.A., F. Bouchot, and A.L. Demain, *Regulation of cellulase formation in Clostridium thermocellum.* J. Gen. Microbiol., 1985. **131**: p. 2303-2308.

118. Hammerstrom, R.A., et al., *The constitutive nature of bacterial cellulases.* Arch. Biochem. Biophys., 1955. **56**(1): p. 123-129.

119. Stevenson, D.M. and P.J. Weimer, *Expression of 17 genes in Clostridium thermocellum ATCC 27405 during fermentation of cellulose or cellobiose in continuous culture.* Appl. Environ. Microb., 2005. **71**(8): p. 4672-4678.

120. Dror, T.W., et al., *Regulation of major cellulosomal endoglucanases of Clostridium thermocellum differs from that of a prominent cellulosomal xylanase.* J. Bacteriol., 2005. **187**(7): p. 2261-2266.

121. Dror, T.W., et al., *Regulation of expression of scaffoldin-related genes in Clostridium thermocellum.* J. Bacteriol., 2003. **185**(17): p. 5109-5116.

122. Kosugi, A., K. Murashima, and R.H. Doi, *Characterization of xylanolytic enzymes in Clostridium cellulovorans: expression of xylanase activity dependent on growth substrates.* J. Bacteriol., 2001. **183**(24): p. 7037-7043.

123. Han, S.O., et al., *Regulation of expression of cellulosomal cellulase and hemicellulase genes in Clostridium cellulovorans.* J. Bacteriol., 2003. **185**(20): p. 6067-6075.

124. Newcomb, M., C.-Y. Chen, and J.H.D. Wu, *Induction of the celC operon of Clostridium thermocellum by laminaribiose.* P. Natl. Acad. Sci. USA, 2007. **104**(10): p. 3747-3752.

125. Ong, S.-E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat. Chem. Biol., 2005. **1**(5): p. 252-262.

126. Fournier, M.L., et al., *Multidimensional separations-based shotgun proteomics.* Chem. Rev., 2007. **107**(8): p. 3654-3686.

127. Yates III, J.R., *Mass spectrometry and the age of the proteome.* J. Mass Spectrom., 1998. **33**(1): p. 1-19.

128. Yates III, J.R., et al., *Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry.* Methods Mol. Biol., 1999. **112**: p. 553-569.

129. Mann, M., R.C. Hendrickson, and A. Pandey, *Analysis of proteins and proteomes by mass spectrometry.* Annu. Rev. Biochem., 2001. **70**(1): p. 437-473.

130. Yates III, J.R., et al., *Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.* Anal. Chem., 1995. **67**(8): p. 1426-1436.

131. Tabb, D.L., J.K. Eng, and J.R. Yates III, *Protein Identification by SEQUEST*, in *Proteome Research: Mass Spectrometry*, P. James, Editor. 2001, Springer. p. 125-142.

132. MacCoss, M.J. and D.E. Matthews, *Quantitative MS for Proteomics: Teaching a New Dog Old Tricks*, in *Anal. Chem. A-Pages*. 2005. p. 294A-302A.

133. Chong, P.K., et al., *Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: implication of multiple injections.* J. Proteome Res., 2006. **5**(5): p. 1232-1240.

134. Heller, M., et al., *Trypsin catalyzed ^{16}O-to-^{18}O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-*

QTOF, and ESI-ion trap mass spectrometers. J. Am. Soc. Mass Spectr., 2003. **14**(7): p. 704-718.

135. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat. Biotechnol., 1999. **17**(10): p. 994-999.

136. Oda, Y., et al., *Accurate quantitation of protein expression and site-specific phosphorylation.* P. Natl. Acad. Sci. USA, 1999. **96**(12): p. 6591-6596.

137. MacCoss, M.J., et al., *A correlation algorithm for the automated quantitative analysis of shotgun proteomics data.* Anal. Chem., 2003. **75**(24): p. 6912-6921.

138. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol. Cell. Proteomics, 2005. **4**(9): p. 1265-1272.

139. Kuster, B., et al., *Scoring proteomes with proteotypic peptide probes.* Nat. Rev. Mol. Cell Biol., 2005. **6**(7): p. 577-583.

140. Bergeron, J.J.M. and M. Hallett, *Peptides you can count on.* Nat Biotech, 2007. **25**(1): p. 61-62.

141. Mallick, P., et al., *Computational prediction of proteotypic peptides for quantitative proteomics.* Nat Biotech, 2007. **25**(1): p. 125-131.

142. Lu, P., et al., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nat. Biotech., 2007. **25**(1): p. 117-124.

143. Zhang, Y. and L.R. Lynd, *Quantification of cell and cellulase mass concentrations during anaerobic cellulose fermentation: development of an*

enzyme-linked immunosorbent assay-based method with application to Clostridium thermocellum batch cultures. Anal. Chem., 2003. **75**(2): p. 219-227.

144. Wright, A., et al., *Proteomic analysis of cell surface proteins from Clostridium difficile.* Proteomics, 2005. **5**(9): p. 2443-2452.

145. Walseth, C.S., *Occurrence of cellulases in enzyme preparations from microorganisms.* TAPPI, 1952. **35**(5): p. 233-238.

146. Morag, E., E.A. Bayer, and R. Lamed, *Affinity digestion for the near-total recovery of purified cellulosome from Clostridium thermocellum.* Enzyme Microb. Tech., 1992. **14**: p. 289-292.

147. Kinter, M. and N.E. Sherman, *Protein Sequencing and Identification Using Tandem Mass Spectrometry.* 2000, New York: John Wiley & Sons, Inc. 301.

148. Emanuelsson, O., et al., *Locating proteins in the cell using TargetP, SignalP and related tools.* Nat. Prot., 2007. **2**(4): p. 953-971.

149. Tabb, D.L., W.H. McDonald, and J.R. Yates, *DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics.* J. Proteome Res., 2002. **1**(1): p. 21-26.

150. Krokhin, O.V., et al., *Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS.* Anal. Chem., 2006. **78**(17): p. 6265-6269.

151. Grishutin, S.G., et al., *Specific xyloglucanases as a new class of polysaccharide-degrading enzymes.* Biochimica et Biophysica Acta (BBA) - General Subjects, 2004. **1674**(3): p. 268-281.

152. Lemaire, M., et al., *OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domains to components of the cell envelope.* J. Bacteriol., 1995. **177**(9): p. 2451-2459.

153. Hayashi, H., et al., *Sequence of xynC and properties of XynC, a major component of the Clostridium thermocellum cellulosome.* J. Bacteriol., 1997. **179**(13): p. 4246-4253.

154. Hazlewood, G.P., et al., *Endoglucanase E, produced at high level in Escherichia coli as a lacZ' fusion protein, is part of the Clostridium thermocellum cellulosome.* Enzyme Microb. Tech., 1990. **12**(9): p. 656-662.

155. Zverlov, V.V., et al., *Multidomain structure and cellulosomal localization of the Clostridium thermocellum cellobiohydrolase CbhA.* J. Bacteriol., 1998. **180**(12): p. 3091-3099.

156. Kurokawa, J., et al., *Clostridium thermocellum cellulase CelT, a family 9 endoglucanase without an Ig-like domain or family 3c carbohydrate-binding module.* Appl. Microbiol. Biot., 2002. **59**(4): p. 455-461.

157. Hayashi, H., et al., *Nucleotide sequences of two contiguous and highly homologous xylanase genes xynA and xynB and characterization of XynA from Clostridium thermocellum.* Appl. Microbiol. Biot., 1999. **V51**(3): p. 348-357.

158. Zverlov, V.V., et al., *Purification and cellulosomal localization of Clostridium thermocellum mixed linkage β-glucanase LicB (1,3–1,4-β-D-glucanase).* Biotechnol. Lett., 1994. **V16**(1): p. 29-34.

159. Halstead, J.R., et al., *A family 26 mannanase produced by Clostridium thermocellum as a component of the cellulosome contains a domain which is*

conserved in mannanases from anaerobic fungi. Microbiology, 1999. **145**(11): p. 3101-3108.

160. Joliff, G., P. Beguin, and J.-P. Aubert, *Nucleotide sequence of the cellulase gene celD encoding endoglucanase D of Clostridium thermocellum.* Nucleic Acids Res., 1986. **14**(21): p. 8605-8612.

161. Zverlov, V.V., K.-P. Fuchs, and W.H. Schwarz, *Chi18A, the endochitinase in the cellulosome of the thermophilic, cellulolytic bacterium Clostridium thermocellum.* Appl. Environ. Microb., 2002. **68**(6): p. 3176-3179.

162. Leibovitz, E., et al., *Characterization and subcellular localization of the Clostridium thermocellum scaffoldin dockerin binding protein SdbA.* J. Bacteriol., 1997. **179**(8): p. 2519-2523.

163. Woodson, K. and K.M. Devine, *Analysis of a ribose transport operon from Bacillus subtilis.* Microbiology, 1994. **140**(8): p. 1829-1838.

164. Gilson, E., et al., *Evidence for high affinity binding-protein dependent transport systems in Gram-positive bacteria and in Mycoplasma.* EMBO J., 1988. **7**: p. 3971-3974.

165. Kruus, K., et al., *Exoglucanase activities of the recombinant Clostridium thermocellum CelS, a major cellulosome component.* J. Bacteriol., 1995. **177**(6): p. 1641-1644.

166. Hon-nami, K., et al., *Separation and characterization of the complexes constituting the cellulolytic enzyme system of Clostridium thermocellum.* Arch. Microbiol., 1986. **145**(1): p. 13-19.

167. Mori, Y., *Comparison of the cellulolytic systems of Clostridium thermocellum YM4 and JW20.* Biotechnol. Lett., 1992. **14**(2): p. 131-136.

168. Wyman, C.E., *Handbook on Bioethanol: Production and Utilization.* 1996: Taylor & Francis. 442.

169. Royer, J.C. and J.P. Nakas, *Interrelationship of xylanase induction and cellulase induction of Trichoderma longibrachiatum.* Appl. Environ. Microb., 1990. **56**(8): p. 2535-2539.

170. Rodriguez, H. and F. Alea, *Regulation of xylanase activity in Cellulomonas sp. IIbc.* Acta Biotechnol., 1992. **12**(4): p. 337-343.

171. Zeilinger, S., et al., *Different inducibility of expression of the two xylanase genes xyn1 and xyn2 in Trichoderma reesei.* J. Biol. Chem., 1996. **271**(41): p. 25624-25629.

172. Shevchenko, A., et al., *In-gel digestion for mass spectrometric characterization of proteins and proteomes.* Nat. Protocols, 2007. **1**(6): p. 2856-2860.

APPENDIX A

In silico classification of proteins from *C. thermocellum* database

Using the 2007/02/16 release of the *C. thermocellum* genome available at NCBI courtesy of DOE, Joint Genome Institute (http://www.ncbi.nlm.nih.gov, Refseq accession number NC_009012), protein sequences annotated to possess a glycoside hydrolase or carbohydrate esterase fold or to participate in the cellulosome were submitted to InterProScan, protein BLAST and SignalP to verify the presence of Doc1, Doc2, Coh1, Coh2 domains, and a signal peptide cleavage site indicating that the protein is secreted from the cell. The presence of a Doc1 indicates that the protein, if it is secreted, would have the ability to bind to CipA and participate in the cellulosome.

Table A1. Checklist for cellulolytic and hemicellulolytic enzymes and structural proteins with or without Doc1, Doc2, Coh1, Coh2 domains, and a signal peptide cleavage site (SignalP) indicating that the protein is secreted from the cell, ranked by GenInfo ID number

Doc1	Doc2	Coh1	Coh2	SignalP	GenInfo ID	Protein name and/or (putative) function and/or modules of interest
✓				✓	125972540	α-L-arabinofuranosidase B, GH43
✓				✓	125972556	GH26
				✓	125972564	CelI β-1,4-cellobiosidase
✓				✓	125972567	CelN endoglucanase (GH9)
✓				✓	125972568	CseP spore coat assembly protein
				✓	125972595	CelY β-1,4-cellobiosidase
✓				✓	125972633	Unknown cellulosome enzyme
✓				✓	125972714	Serine protease inhibitor I4, serpin
✓				✓	125972735	LicB lichenase (GH16)
✓				✓	125972761	Unknown cellulosome enzyme
✓				✓	125972768	Integrins α chain, CBD6
✓				✓	125972780	Unknown cellulosome enzyme
✓				✓	125972791	CelA endoglucanase (GH8)
✓				✓	125972792	ChiA chitinase (GH18)
✓				✓	125972796	GH9
✓				✓	125972926	GH5
✓				✓	125972933	CelK cellobiohydrolase (GH9)
✓				✓	125972934	CbhA cellobiohydrolase (GH9)
✓				✓	125972954	GH9
✓				✓	125972956	Unknown cellulosome enzyme
✓				✓	125972959	Unknown cellulosome enzyme
			✓	✓	125972973	Cellulosome anchor protein
✓				✓	125973055	CelB endoglucanase (GH5)
✓				✓	125973062	CelF endoglucanase (GH9)
✓				✓	125973097	CelR endoglucanase (GH9)
✓				✓	125973142	CelJ endoglucanase (GH9), Ig-like
✓				✓	125973143	CelQ endoglucanase (GH9)
✓				✓	125973158	Endopygalactorunase
✓				✓	125973178	GH81
✓				✓	125973179	Galactan β-1,3-galactosidase (GH43): ricin B lectin
✓				✓	125973247	Unknown cellulosome enzyme
			✓	✓	125973253	Cellulosome anchor protein
			✓	✓	125973254	Cellulosome anchor protein
✓				✓	125973263	GH9
✓				✓	125973315	CelE endoglucanase (GH5), CE2
✓				✓	125973316	Lipase GDSL
✓				✓	125973339	GH5: coagulation factor 5/8 type-like
✓				✓	125973343	CelD endoglucanase (GH9)
✓				✓	125973429	XynY xylanase (GH10), CE1
✓				✓	125973435	Unknown cellulosome enzyme
					125973750	β-1,4-cellobiosidase, SLH, Ig-like fold, Fn type III-like fold
✓				✓	125973786	GH43, CBD6
			✓	✓	125973822	SdbA cell-surface anchor protein
✓				✓	125973912	XghA xyloglucanase (GH74)
✓				✓	125973914	Arabinogalactan endo-1,4-galactosidase (GH53)
					125973983	GH5
✓				✓	125973984	CelH endoglucanase (GH5), GH26, CBD11
✓				✓	125974310	Unknown cellulosome enzyme
✓				✓	125974342	XynC xylanase (GH10)
✓				✓	125974394	Unknown cellulosome enzyme
✓				✓	125974464	XynZ xylanase (GH10), CE1

Table A1. (continued)

Doc1	Doc2	Coh1	Coh2	SignalP	GenInfo ID	Protein name and/or (putative) function and/or modules of interest
✓				✓	125974530	Unknown cellulosome enzyme
✓				✓	125974579	CelS cellobiohydrolase (GH48)
					125974609	GH10
✓				✓	125974624	Unknown cellulosome enzyme
✓					125974625	GH43, α-L-arabinofuranosidase B
✓				✓	125974626	α-L-arabinofuranosidase B, GH30
				✓	125974652	GH8
✓				✓	125974678	GH5, CBD6
✓				✓	125974679	GH10, CBD4, CBD6
✓				✓	125974680	CBD4, CBD6, pectin lyase-fold
✓				✓	125974681	GH43, CBD4, CBD6
✓					125974682	GH2, GH2, GH2, Ig-like, CBD4, CBD6
✓				✓	125974756	Unknown cellulosome enzyme
✓					125974845	GH9, CBD3a
					125975032	α-N-arabinofuranosidase
✓					125975033	Unknown
✓				✓	125975073	XynD xylanase (GH10)
✓				✓	125975242	GH9, CBD3a
✓				✓	125975243	GH9, CDB3a
					125975289	CelC (GH5)
				✓	125975291	LicA (GH16), SLH domain, CBD CenC-like
✓				✓	125975294	CelT endoglucanase (GH9)
✓				✓	125975353	CelG endoglucanase (GH5)
✓				✓	125975360	Unknown cellulosome enzyme
				✓	125975376	Chitinase (GH18)
✓				✓	125975452	XynA xylanase (GH11), acetylxylan esterase (CE4)
✓				✓	125975491	GH30, CBD6
	✓	✓		✓	125975556	CipA scaffolding protein
			✓	✓	125975557	OlpB cell-surface anchor protein
			✓	✓	125975558	Orf2p cell-surface anchor protein
		✓		✓	125975559	OlpA cellulosome anchor protein
✓				✓	125975610	Unknown cellulosome enzyme
✓				✓	125975619	Rhamnogalacturan acetylesterase-like, lipolytic enzyme G-D-S-L

APPENDIX B

Freeze-down procedure for culture collection

Some laboratories keep frozen stocks of *C. thermocellum* spores. A liquid *C. thermocellum* culture is left at room temperature for several days (at least a week) and then frozen at -80°C. This procedure does not require the use of glycerol. Revival of spores can involve a lag of 2-3 days.

We have preferred to freeze growing cultures of *C. thermocellum*, which can be revived in 1 day. A solution containing 40% (vol/vol) glycerol, 0.5 g/L L-cysteine HCl, and 0.0001% (wt/vol) resazurin is prepared and transferred to an anaerobic culture bottle. The solution is then simultaneously heated and sparged with nitrogen gas for 5-10 min, before the bottle is sealed with a rubber stopper and aluminum cap and autoclaved. Meanwhile, clean empty 10-mL anaerobic culture bottles are flushed with nitrogen gas, stoppered and autoclaved. When the glycerol solution emerges from the autoclave, it should be clear in colour. Inside the anaerobic chamber, 5-mL volumes of the sterile glycerol solution is transferred to the sterile 10-mL bottles using a syringe. Because of the viscosity of glycerol, even at 40%, transfer by syringe is easier if the solutions are heated in the anaerobic chamber's incubator. If the glycerol solution changes colour (to pink or orange) during transfer, the bottle should not be used.

Liquid cultures are grown to log phase. The growth of a cellobiose-grown culture can be easily monitored by measuring the optical density (OD) at 660 nm (Figure B1). Exponential phase is reached at an OD of 0.3-0.5. Inside the anaerobic chamber, 5-mL

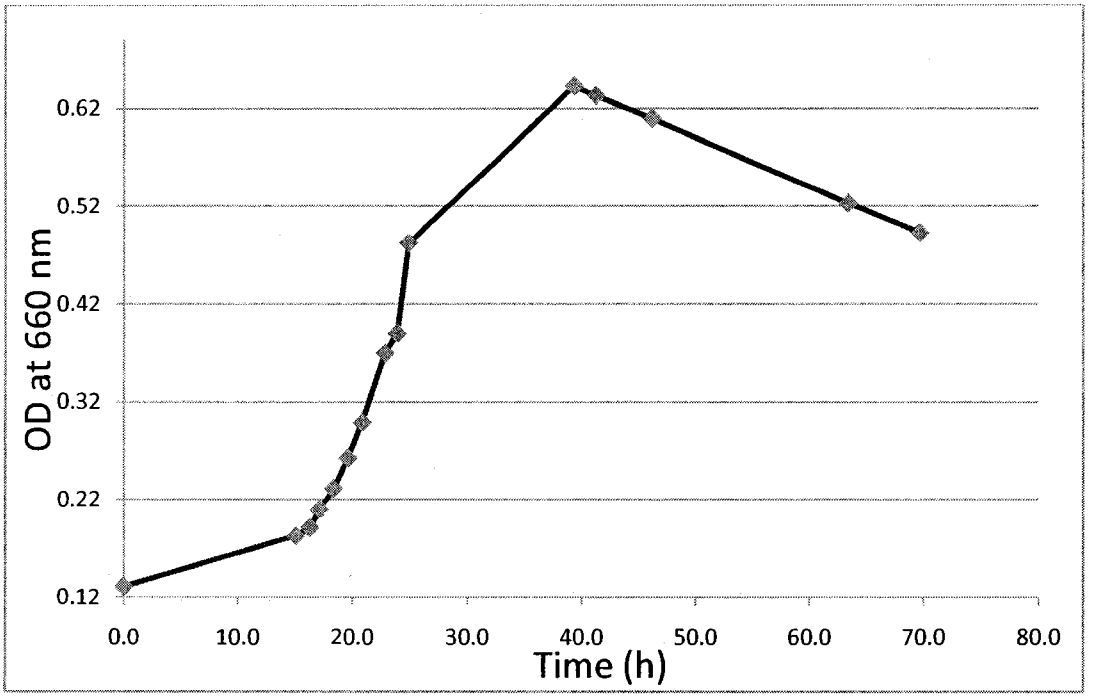

Figure B1. Growth of C. thermocellum on 0.5% (wt/vol) cellobiose from 10% (vol/vol) inoculum from Avicel-grown culture in exponential phase

aliquots of the log-phase culture are transferred by syringe to the 10-mL culture bottles. These are then frozen at -80°C.

APPENDIX C

In-gel trypsin digestion protocol

The following protocol is adapted from protocols such as found in [147, 172]. The procedure requires a period of 3-4 days. The volumes noted are for gel bands of less than 20μL. For a 1D gel, this corresponds to a band with dimensions 1-2 mm × 1cm × 1 mm (band width × lane width × gel thickness). For gel bands larger in size, volumes of each reagent should be increased proportionally. All reagents must be prepared fresh. The trypsin used should be proteomics grade and modified to resist autolysis and be defective for chymotrypsin-activity (such as T6567 from Sigma-Aldrich). Solvents should be HPLC grade. Water for preparing solutions should be HPLC-grade or nanopure or at worst double-distilled. Prior to the addition of trypsin, the same pipette (tip) can be used for removal of reagent from sample all tubes.

Coomassie destaining solution

50% (vol/vol) methanol and 10% (vol/vol) acetic acid

Buffers

100 mM ammonium bicarbonate: 0.0791 g of ammonium bicarbonate per 10 mL of water

50 mM ammonium bicarbonate: ½ dilution of above

Reductant

10 mM DTT: 15 mg of DTT per 10 mL of buffer

Keep in dark.

Alkylating agent

100 mM iodoacetic acid: 18 mg of iodoacetic acid per mL of buffer

Keep in dark.

Trypsin

0.2 ng/μL trypsin: 20 μg in 1 mL ice-cold buffer

Prepare just prior to use and keep on ice.

Dehydration or extraction solvents

ACN

50/5% (vol/vol) ACN/FA

80/5% (vol/vol) ACN/FA

Day 1: Excision and destaining of protein gel bands

1. Using a scalpel or razor blade, cut the protein bands from the gel as closely as possible, then divide each gel band into smaller pieces, approximately 1-2 mm^3 in size. Be careful not to crush or the gel pieces, which could result in fine particles that can block your LC system.
2. Place the gel pieces for each band into a 1.5-mL microcentrifuge tube.
3. Add 200 μL of the destaining solution and allow gel pieces to destain for 4 h or overnight.

Day 2: Reduction, alkylation, digestion of protein with trypsin

4. Carefully remove the destaining solution from the sample. Repeat steps 3-4 as necessary.
5. Add 200 µL of ACN, vortex mix once, and let sit for 5 min to dehydrate at room temperature. When dehydrated, the gel pieces will have an opaque white color and will be significantly smaller in size.
6. Carefully remove the ACN from the sample and discard.
7. Completely dry the gel pieces at room temperature in a vacuum centrifuge for 2-3 min.
8. Add 30 µL of 10 mM DTT and reduce the protein for 30 min at room temperature.
9. Carefully remove the DTT solution from the sample with a plastic pipette and discard.
10. Add 30 µL of 100 mM iodoacetic acid to alkylate the protein at room temperature for 30 min.
11. Carefully remove the iodoacetic acid solution from the sample and discard.
12. Add 200 µL of ACN and dehydrate the gel pieces for ~5 min at room temperature. When dehydrated, the gel pieces will have an opaque white color and will be significantly smaller in size.
13. Carefully remove the ACN from the sample with a plastic pipette and discard.
14. Rehydrate the gel pieces in 200 µL of 100 mM ammonium bicarbonate, incubating the samples for 10 min at room temperature.
15. Carefully remove the ammonium bicarbonate from the sample and discard.

16. Add 200 µL of ACN, vortex mix once, and dehydrate the gel pieces for 5 min at room temperature.
17. Carefully remove the ACN from the sample and discard.
18. Completely dry the gel pieces at ambient temperature in a vacuum centrifuge for 2-3 min.
19. Prepare the trypsin reagent as above.
20. Add 30 µL of the trypsin solution to the sample and allow the gel pieces to rehydrate on ice for 10 min with occasional vortex mixing. Watch that the gel pieces appear to have been rehydrated by the trypsin solution.
21. Drive the gel pieces to the bottom of the tube by centrifuging the sample for 30 s. Carefully remove the excess trypsin solution from the sample and discard.

Day 3: Extraction of peptides

22. Add 5 µL of 50 mM ammonium bicarbonate to the sample (or enough to cover the gel pieces). Vortex mix once. Drive the sample to the bottom of the tube by centrifuging for 30 s, and carry out the digestion for 18 h or overnight at 37°C.
23. Add 30 µL of 50 mM ammonium bicarbonate to the digest and incubate the sample for 10 min with occasional gentle vortex mixing. Drive the digest to the bottom of the tube by centrifuging the sample for 30 s. Carefully collect the supernatant and transfer the sample to a new microcentrifuge tube.
24. Add 30 µL of the 50/5% (vol/vol) ACN/FA solution to the tube containing the gel pieces, and incubate the sample for 10 min with occasional gentle vortex mixing. Drive the extract to the bottom of the tube by centrifuging the sample for 30 s.

Carefully collect the supernatant and combine the extract in the same new microcentrifuge tube.

25. Repeat step 24.
26. Add 30 μL of the 80/5% (vol/vol) ACN/FA solution to the tube containing the gel pieces, and incubate the sample for 10 min with occasional gentle vortex mixing. Drive the extract to the bottom of the tube by centrifuging the sample for 30 s. Carefully collect the supernatant and combine the extract in the same new microcentrifuge tube.
27. Reduce the volume of the extract to less than 20 μL by evaporation in a vacuum centrifuge at ambient temperature. Do not allow the extract to dry completely.
28. Adjust the volume of the digest to 10-20 μL, as needed, with 5/0.1% (vol/vol) ACN/FA for analysis.

APPENDIX D

In-solution trypsin digestion for cellulosomal protein

Cellulosomes are notoriously difficult to dissociate. Dissociation of the complex is crucial to in-solution trypsin digestion of its components. Some studies have reported the need to use SDS, EDTA plus DTT to break the complex into its component parts [68]; others have used SDS and temperatures of 70°C [84]. Trypsin can tolerate up to 0.1% SDS (wt/vol). The ideal protease/protein ratio is between 1/100 and 1/50 (wt/wt). All reagents must be prepared fresh. The trypsin used should be proteomics grade and modified to resist autolysis and be defective for chymotrypsin-activity (such as T6567 from Sigma-Aldrich). Solvents should be HPLC grade. Water for preparing solutions should be HPLC-grade or nanopure or at worst double-distilled.

Buffer

100 mM ammonium bicarbonate, pH 7.8: 0.0791 g per 10 mL of water

Reductant

200 mM DTT: 30 mg of DTT per mL of buffer

Keep in dark.

Alkylating agent

200 mM iodoacetic acid: 36 mg of iodoacetic acid per mL of buffer

Keep in dark.

Detergent

10% SDS (wt/vol)

Trypsin

0.2 mg/mL trypsin: 20 μg in 100 μL of ice-cold buffer

Prepare just prior to use and keep on ice.

1. Dry down approximately 200 μg of sample protein in a vacuum centrifuge at medium dry rate. (For example, if your protein concentration is 1 mg/mL, dry down 200 μL.)
2. Suspend the dehydrated protein in 100 μL of 100 mM ammonium bicarbonate.
3. Add 4 μL of 10% (wt/vol) SDS* and 5 μL of 200 mM DTT (for final concentration of 0.37% SDS and 18 mM DTT). Transfer to a 0.5-mL PCR tube.
4. Incubate at 70°C for 45 min in a thermocycler with heated lid.
5. Add 20 μL of 200 mM iodoacetic acid. Incubate at room temperature for 1 h in the dark.
6. Add 20 μL of DTT (to quench the alkylation reaction). Incubate at room temperature for 1 h in the dark.
7. Add 350 μL of 100 mM ammonium bicarbonate (to dilute the SDS concentration to 0.08%).*
8. Add 10 μL of trypsin to the alkylated protein suspension. Incubate 4-8 h or overnight at 37°C.

9. Place the proteolysis reaction at 4°C. Remove a 30-µL aliquot to test the extent of the digestion by SDS-PAGE.
10. If the reaction is complete (no proteins visible on gel), then the reaction can be stopped by adding 20 µL of glacial acetic acid. If not, repeat steps 8-9.
11. Reduce the volume to less than 20 µL by evaporation in a vacuum centrifuge at medium dry rate. Do not allow the extract to dry completely.

* If this protocol does not yield a total protein digest, consider increasing the amount of SDS added (step 3), and then increasing the dilution accordingly (step 7).

APPENDIX E

Attempts to calibrate the emPAI method

Ishihama *et al.* showed an almost linear relationship between protein concentration and 10^{PAI} - 1, where PAI (protein abundance index) is equal to the ratio of the number of unique peptides observed for a protein over the number of theoretically observable peptides for that protein [138]. Their results for 46 protein digests are reproduced in Figure E1. We have endeavoured to establish a direct relationship between concentration and emPAI, using our in-house nanoLC-ESI-MS conditions. Using the same chromatographic and MS conditions as described in section 2.4, emPAI values were determined for triplicate 2-μL injections of a series of 5 dilutions of a 3-protein digest mixture (Table E1). The digest mixture was prepared by combining digest standards yeast protein and bovine serum albumin from Michrom Bioresources. A linear regression with R^2 of 0.69 was the best fit for a plot emPAI values versus protein digest concentrations (Figure E2), even though different regression types were tried. The curve seems to flatten out at higher concentrations such that proteins in higher abundance would be underestimated. While the emPAI method has thus been shown to provide only semi-quantitative information, it is nonetheless useful as a means of ranking protein abundances within a sample.

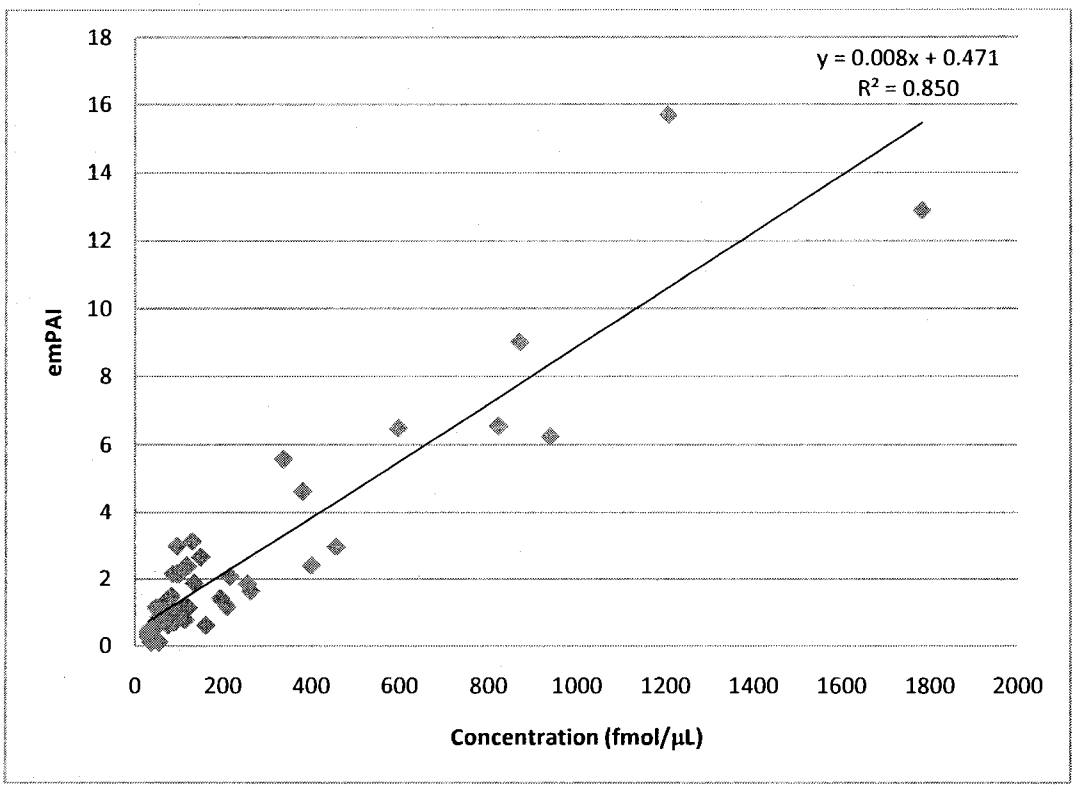

Figure E1. Calibration of emPAI method by Ishihama *et al.* [138]. A plot of emPAI values for 46 protein digests against the concentrations at which they were analyzed by LC-MS yielded a linear regression with R^2 of 0.85.

Table E1. EmPAI values for 2-μL injections of a mixture of 3 protein digests at varying concentrations

Protein	no. obs. un. pep.	emPAI	Conc. (fmol/μL)	Injection no.
Yeast enolase 1	4	0.33	3.13	1
(gi 119336)	4	0.33	3.13	2
	4	0.33	3.13	3
	7	0.65	6.25	4
	6	0.54	6.25	5
	6	0.54	6.25	6
	11	1.21	12.5	7
	11	1.21	12.5	8
	10	1.05	12.5	9
	12	1.37	25.0	10
	10	1.05	25.0	11
	8	0.78	25.0	12
	13	1.55	50.0	13
	9	0.91	50.0	14
	10	1.05	50.0	15
Bovine serum albumin	13	0.66	31.3	1
(gi 1351907)	16	0.87	31.3	2
	12	0.60	31.3	3
	15	0.80	62.5	4
	17	0.94	62.5	5
	18	1.02	62.5	6
	22	1.36	125	7
	20	1.18	125	8
	17	0.94	125	9
	29	2.10	250	10
	25	1.65	250	11
	25	1.65	250	12
	28	1.98	500	13
	32	2.49	500	14
	26	1.76	500	15
Yeast alcohol dehydrogenase I	6	0.74	31.3	1
(gi 1168350)	8	1.09	31.3	2
	8	1.09	31.3	3
	8	1.09	62.5	4
	7	0.91	62.5	5
	6	0.74	62.5	6
	8	1.09	125	7
	11	1.75	125	8
	9	1.29	125	9
	14	2.63	250	10
	11	1.75	250	11
	14	2.63	250	12
	12	2.02	500	13
	14	2.63	500	14
	15	2.98	500	15

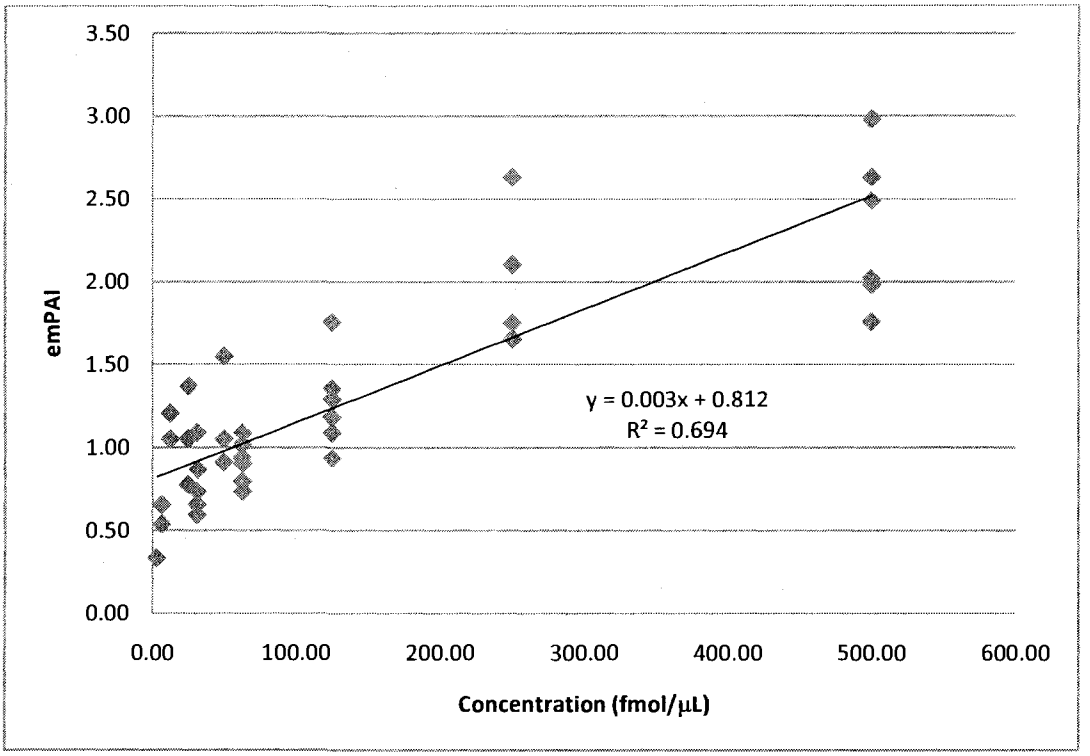

Figure E2. In-house calibration of emPAI method. A plot of emPAI values for 3 protein digests against the concentrations at which they were analyzed by nanoLC-ESI-MS yielded a linear regression with R^2 of 0.69.

APPENDIX F

RelEx procedure using BioWorks 3.3 and DTASelect 1.9

For RelEx analysis, the .DTA file-containing .RAW file is placed in its own specific folder. SEQUEST is then used to search the .RAW file, generating .OUT file results in the same folder, rather than a unified .SRF file. A sequest.params file referring to the appropriate non-indexed .FASTA file is created, once again, in that same folder. It is important that the file be called 'sequest.params' and that the original non-indexed .FASTA file be used (Figure F1). Furthermore, since the non-indexed .FASTA file is being used, the PTMs will have to be entered into the search parameters for searching dynamically on the fly (for δ masses see section 2.5 on Database screening and success criteria).

Once the search is complete, open a DOS command prompt. Navigate to the same folder as above and type 'dtaselect --here' (Figure F2).

When DTASelect has completed its protein calling, open RelEx. Click on Tools in the navigation bar and then Extract-Chro (Figure F3). Navigate to the same folder as above and choose the DTASelect-filter.txt file that was created in the folder. Change the Atomic Enrichment of Label to 99%, and click on Extract. When the extraction is complete, close the Extract-Chro window and click on Options in the navigation bar and then Integration Settings. Make sure all of the check-boxes are checked as shown in Figure F3. Under Chromatogram Filters, change the minimum correlation factors at 1 and 10 as desired; the higher they are, the less manual filtering will be necessary: values of 0.9 were used for the work described here. Under Protein Filters, change the minimum

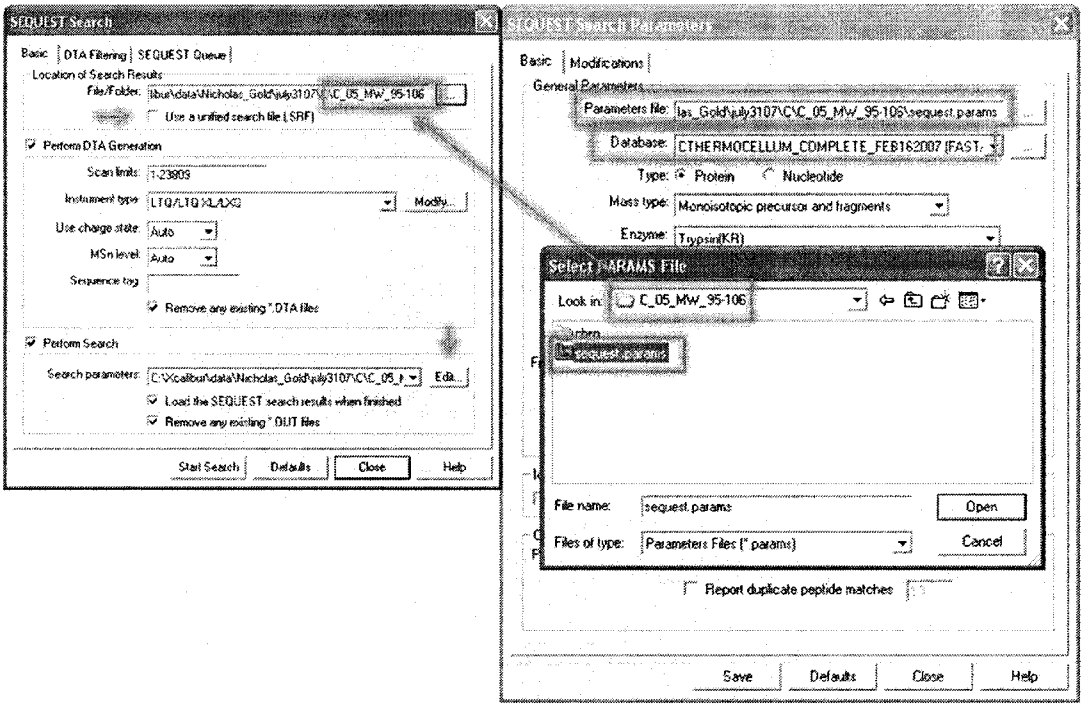

Figure F1. BioWorks 3.1 SEQUEST search parameters for RelEx analysis. The .SRF option must be unchecked. The folder for the search results must match the folder for the sequest.params file, which must point to an indexed .FASTA sequence database.

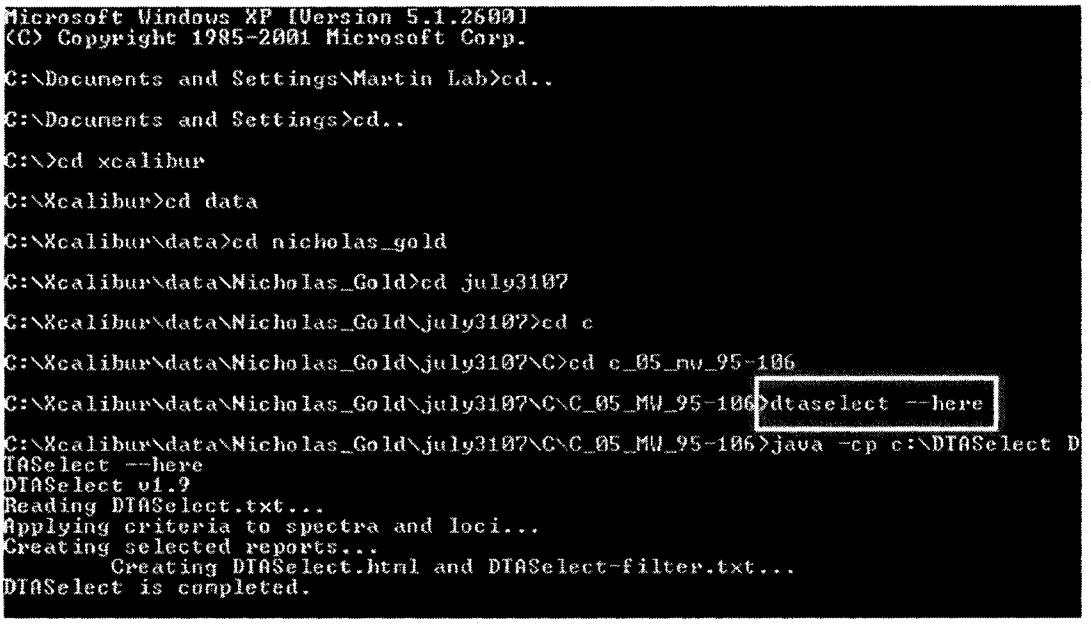

Figure F2. DTASelect deployment via DOS command prompt. Type 'dtaselect --here' once at the appropriate folder.

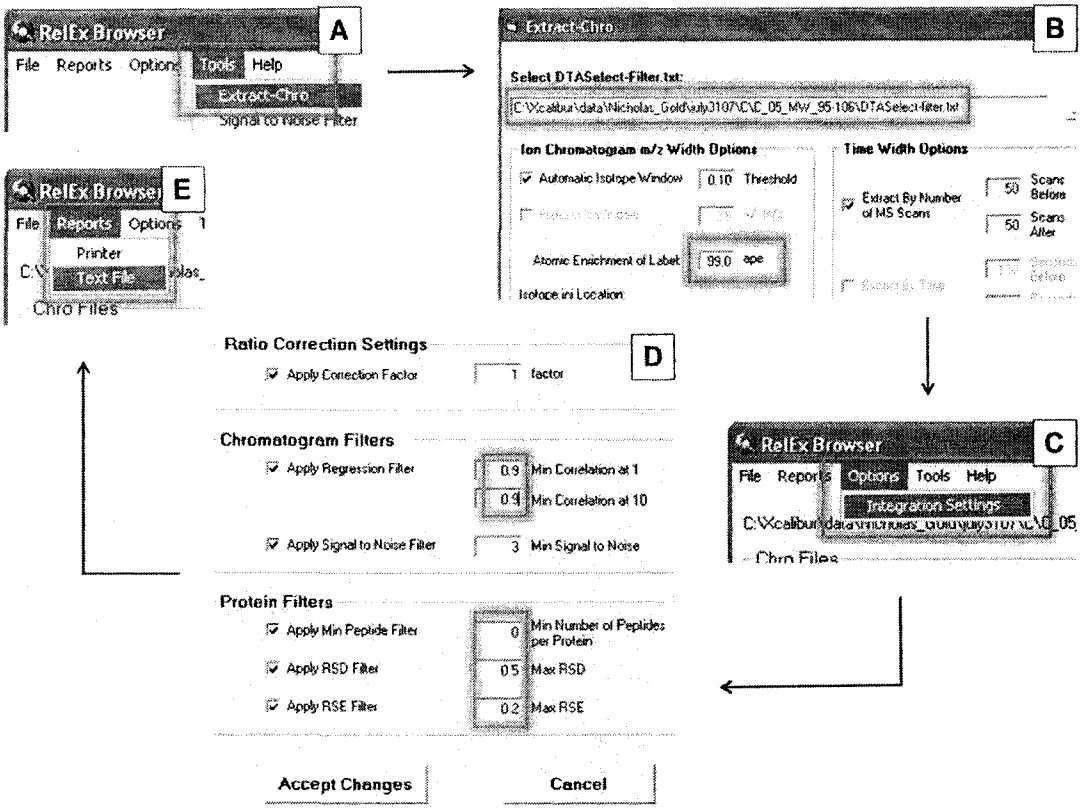

Figure F3. Steps for analysis of peptide ratios in RelEx Browser. (a) Open Extract-Chro tool. (b) Navigate to DTASelect-filter.txt in appropriate folder and set Atomic Enrichment of Label to 99%. (c) Open Integration Settings Option. (d) Check all boxes for Ratio Correction Settings, Chromatogram Filters and Protein Filters. Set Min Correlation factors to 0.9. Set Min Number of Peptides to 0. (e) Create a report by clicking on Text File.

number of peptides per protein to 0, so that even if only 1 peptide ratio was calculated for a given protein, it will be reported (the default is value is 2); single ratios of the kind can be combined with other ratios calculated for the same protein should they appear in different .RAW files. Finally, to generate a report, click Report in the navigation bar, followed by Text. A tab-delineated Protein-Output.txt file will appear in a folder name 'chro', under the same folder as above.